# Adaptive RD Optimized Hybrid Sound Coding

H. van Schijndel, Nicolle; Bensa, Julien; Christensen, Mads Græsbøll; Colomes, Catherine; Edler, Bernd; Heusdens, Richard; Jensen, Jesper; Jensen, Søren Holdt; Kleijn, W. Bastiaan; Kot, Valery; Kovesi, Bala Zs; Lindblom, Jonas; Massaloux, Dominique; A. Niamut, Omar; Nordén, Fredrik; H. Plasberg, Jan; Vafin, Renat; Van De Par, Steven; Virette, David; Wûbbolt, Oliver

*Published in:*
Journal of the Audio Engineering Society

*Publication date:*
2008

*Document Version*
Publisher's PDF, also known as Version of record

# Adaptive RD Optimized Hybrid Sound Coding*

**NICOLLE H. van SCHIJNDEL,[1] *AES Member,* JULIEN BENSA,[2] MADS G. CHRISTENSEN,[3] CATHERINE COLOMES,[2] BERND EDLER,[4] *AES Member,* RICHARD HEUSDENS,[5] JESPER JENSEN,[5] SØREN HOLDT JENSEN,[3] W. BASTIAAN KLEIJN,[6] VALERY KOT,[1] BALÁZS KÖVESI,[2] JONAS LINDBLOM,[6] DOMINIQUE MASSALOUX,[2] OMAR A. NIAMUT,[5] *AES Associate Member,* FREDRIK NORDÉN,[3] JAN H. PLASBERG,[6] RENAT VAFIN,[6] STEVEN VAN DE PAR,[1] *AES Member,* DAVID VIRETTE,[2] AND OLIVER WÜBBOLT[4]**

[1]*Philips Research Laboratories, Digital Signal Processing Group, 5656 AE Eindhoven, The Netherlands*
({Nicolle.van.Schijndel, Steven.van.de.Par., Valery.Kot}@philips.com)

[2]*France Télécom/Orange Research Laboratories, F-22307 Lannion, France*
({Catherine.Colomes, Balazs.Kovesi, Dominique.Massaloux, David.Virette}@orange-ftgroup.com, Julien.Bensa@gmail.com)

[3]*Aalborg University, Department of Electronic Systems, DK-9220 Aalborg, Denmark*
({mgc, shj}@es.aau.dk, Fredrik.Norden@ericsson.com)

[4]*University of Hannover, Laboratorium für Informationstechnologie, D-30167 Hannover, Germany*
(edler@tnt.uni-hannover.de, Oliver.Wuebbolt@thomson.net)

[5]*Delft University of Technology, Department of Mediamatics, 2628 CD Delft, The Netherlands*
(R.Heusdens@tudelft.nl, jsj@oticon.dk, Omar.Niamut@tno.nl)

[6]*KTH (Royal Institute of Technology), School of Electrical Engineering, SE10044 Stockholm, Sweden*
({Bastiaan.Kleijn, Jan.Plasberg}@ee.kth.se, {Renat.Vafin, Jonas.Lindblom}@gmail.com)

Traditionally, sound codecs have been developed with a particular application in mind, their performance being optimized for specific types of input signals, such as speech or audio (music), and application constraints, such as low bit rate, high quality, or low delay. There is, however, an increasing need for more generic sound codecs, created by the emergence of heterogeneous networks and the convergence of communication and entertainment devices. To obtain such versatility, this study employs hybrid sound coding based on operational rate-distortion (RD) optimization principles. Applying this concept, a prototype coder has been implemented with emphasis on (dynamic) adaptation to the input and to application constraints. With this prototype, listening tests have been performed for different application scenarios. The results demonstrate the versatility of the concept while keeping competitive sound quality compared to dedicated state-of-the-art codecs.

## 0 INTRODUCTION

Many sound codecs currently exist, based on different coding techniques. Each codec has its own strength and is dedicated to a specific type of input signal and to constraints such as bit rate or delay. For coding of audio (music) signals, general methods include transform coding (MPEG-1/2 layer I, II, III (MP3), MPEG-2/4 AAC [1]) for high bit rates and parametric (sinusoidal) coding ([2], MPEG-4 HILN [3], MPEG-4 SSC [4]) for low rates. For coding speech signals, predictive coding techniques are generally used, such as the GSM-EFR codec, which is widely used in our mobile phones today, and the 3GPP AMR-WB codec [5]. However powerful within their designated application areas, all these codecs are much less efficient for other types of signals and bit rates. As a result, the application flexibility of an individual codec is limited.

The emergence of time-varying heterogeneous networks and the convergence of traditional consumer electronics with mobile communications have created, how-

*Manuscript received 2007 December 20; revised 2008 September 25.

ever, a growing need for more flexible and adaptive—preferably even generic—codecs, and the key question in the fields of speech and audio coding is whether generic coding technology is feasible.

Such generic or, in other words, universal coding is able to adapt seamlessly and in real time to application-imposed constraints and (time-varying) network-imposed constraints on coding attributes, such as bit rate, quality, and latency, and also to the time-varying characteristics of the input signal and to user preferences. As a result of its versatility, a generic codec can be applied in a broad range of applications, such as Internet radio, solid-state audio playback, and mobile communications, without the risk of a mismatch between codec and application.

First steps toward generic coding have been taken in the creation of the MPEG-4 audio standard (see [6]). This standard essentially consists of a large set of codecs, each designed for a specific set of conditions. Switching between these codecs is based on user settings. However, this is not equivalent to true versatility, because the user or system designer is burdened with the task of selecting the right codec for a particular application.

Also hybrid codecs, such as MPEG-4 HE-AAC [7], 3GPP AMR-WB+ [8], [9], ITU-T G.729.1 [10], MPEG-4 CELP+AAC [11], MTPC [12], and SSC+RPE [13], can be interpreted as first steps toward universality, since they exploit the possibility of simultaneous use of several coding techniques. For example, in SSC+RPE a sinusoidal coder, working at 24 kbit/s, captures the tonal parts of the signal; what is left from the target bit rate is used by the predictive coder, based on regular pulse excitation (RPE), to code the residual signal. Since a single coding technique that works well for all bit rates and input signals has not been found, combining the strengths of several techniques is a promising approach for generic coding. However, the key question—how to combine the coding techniques properly—is still open. There is a need for a generic application methodology instead of heuristic methods, which do not suffice for generic coding, because such a methodology should be able to handle changing input signals and constraints, ensuring that for every input signal as well as for every application, network, or user constraint the best possible sound quality or, equivalently, the lowest perceived distortion is attained.

Such demands are in principle inherently met if operational rate-distortion (RD) optimization techniques are employed where a perceptually relevant measure of distortion is used. Building on classical RD optimization [14], these techniques ensure optimal and achievable performance within a given practical coding framework. Automatically setting coder parameters like bit allocation (rate) such that the resulting distortion is minimal, such techniques allow for coding schemes that can, in principle, dynamically adapt to source characteristics and bit rate or other constraints, thereby ensuring excellent coding performance. Apart from flexibility, operational RD optimization methods allow to balance bit-rate spending in different aspects of the signal representation optimally. In this way, in principle, the best solution for distributing bits is found given certain constraints on, for example, the bit rate. The use of a perceptual distortion measure, which predicts the audibility of the signal distortion that is introduced by the encoder, is essential for achieving full benefit from the RD optimization techniques for the listener.

Operational RD optimization was introduced by Shoham and Gersho in their paper on efficient bit allocation [15] and was applied to image compression by Ramchandran and Vetterli [16]. For efficient sound modeling, an approach based on operational RD optimization is described by Prandoni and coworkers [17]–[19] in the context of sinusoidal and predictive coding. Prandoni used a measure based on simple signal characteristics: the mean squared error of the approximated signal. However, since it is the quality as perceived by the listener that counts in the end, the distortion measure should have perceptual relevance. Heusdens and van de Par [20] introduced a perceptually relevant distortion measure in operational RD optimized sinusoidal coding. Van de Par and Kohlrausch [21] confirmed that operational RD optimization leads to better results than the conventional approach where perceptual distortions are distributed equally over time and frequency. The conventional rate and distortion loops in the quantization stage of AAC aim to keep the quantization noise for each frequency band and time instance below the masking threshold. In contrast, with operational RD optimization the encoder can encode a difficult-to-encode segment with relatively poor quality to have extra bits available to create a larger quality improvement in other segments that are easier to encode. The overall perceived quality improves as a result of such optimization. Recently this approach has been described in the context of AAC [22].

This study[1] extends the use of RD optimization to hybrid coding. By combining the strengths of several coding techniques using perceptual operational RD optimization mechanisms to ensure flexibility and adaptivity, a broad range of applications comes within reach, such as broadcasting, storage, and communication. Accordingly the aim of this study is to investigate the feasibility of an RD optimization framework for generic sound coding. Basically the complete coding framework is controlled by RD optimization. Besides applying this for a combination of several coding techniques, it is used for segmentation and for the selection of internal coder settings, such as model selection, component selection, and bit allocation. Apart from flexibility and efficiency, this approach has the advantage that there is no need for specific tuning as is the case for most state-of-the-art codecs.

Our adaptive RD optimized hybrid sound coding framework is described in Section 1. For validation of the concept, listening tests are presented in Section 2 using a prototype coder which incorporates the essential aspects of the framework. This paper ends (Section 3) with a discus-

[1]This paper reports the results of the E.U.-funded project ARDOR (Adaptive RD Optimized sound codeR). For more information, see also the project's website: http://www.hitech-projects.com/euprojects/ardor.

sion and conclusion about the feasibility of using operational RD optimization for generic sound coding.

## 1 FRAMEWORK

### 1.1 Introduction

The basic idea of this study is to employ operational RD optimization techniques, using advanced perceptual distortion measures (explained in Section 1.4) consistently throughout the coding framework. An operational RD control combines the strengths of several coding techniques and adapts to the time-varying input signal, taking into account constraints such as bit rate, which can also be time-varying. To that end the control uses an advanced perceptual distortion measure. Three different coding techniques have been investigated in the framework: sinusoidal coding, transform coding, and CELP coding. (For an overview of these methods, see [23].) These coders are working in cascade, on the residue of their predecessor (multistage coding). In Fig. 1 this is illustrated for the combination of sinusoidal coding followed by transform coding. The control determines how bits are best distributed among the coding techniques on the basis of a perceptual distortion measure. More information about this bit-distribution mechanism is given in Section 1.2.

Operational RD optimization principles are also used to control the time segmentation and bit allocation per segment, which can be different for the different coding techniques. This is illustrated in Fig. 2, which shows the contents of the coding blocks, that is, those saying "sinusoidal" and "transform," of Fig. 1. Fig. 2 shows that the situation is more complex than suggested in Fig. 1, because in between the hybrid RD optimization and the coding techniques, there is another RD optimization layer that controls the segmentation and bit distribution over segments. More information about this can be found in Section 1.3.

Operational RD principles also determine the internal settings of the coding techniques, such as parameter quantization, as illustrated in Fig. 3. So this is the third layer of RD optimization in the framework. For efficient quantization, the masking curve is used. This curve is derived from the excitation pattern, which is a perceptually meaningful representation of the spectral energy distribution of the input signal (see [24]). The excitation pattern is determined by the excitation pattern coder based on the original input signal and available to the other coding techniques, which is illustrated in Fig. 4 for the combination of sinusoidal and transform coding. Furthermore, the excitation pattern information is put in the bit stream and used in the decoder to complete missing parts in the spectrum in case the coding techniques together are not able to model the input signal totally, for instance, at low bit rates. In such cases the decoder adds spectrally shaped noise, consequently retaining the timbre of the original input. More information about the coding techniques is given in Section 1.5. Section 1.6 addresses how the framework can handle stereo input signals.

A full optimization of all encoding parameters that influence rate and distortion is bound to lead to high com-

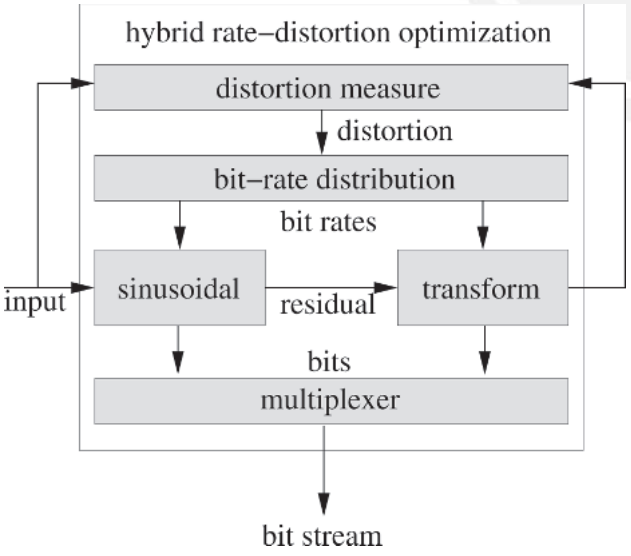



Fig. 1. Operational RD optimized hybrid coding framework.

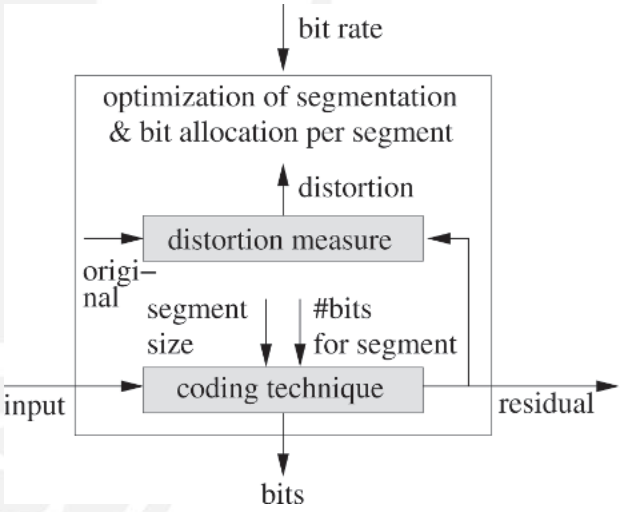



Fig. 2. Segmentation and bit allocation over segments.

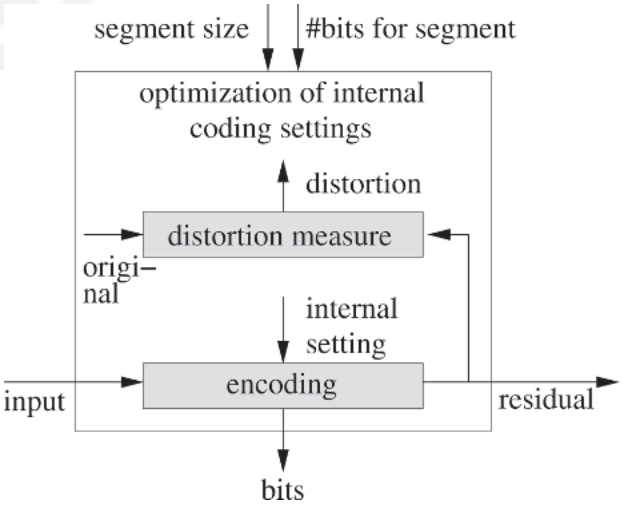



Fig. 3. Internal RD optimization of coding techniques.

putational complexity, and methods for a reduction of complexity are discussed in Section 1.7.

Whereas the foregoing is giving a broad outline of the operational RD optimized hybrid coding framework, more explanation is given in the following subsections. Deliberately the term framework has been used and not coder, because the aim of this study is to examine the concept, more specifically its feasibility for generic sound coding, and for that several aspects of the technology have been investigated separately. Nevertheless, much of the framework has been integrated in a complete prototype coder, which has been evaluated in listening tests, as will be shown toward the end of this paper.

### 1.2 Distribution of Bit Rate among the Coding Techniques and Framing

To combine the strengths of the coding techniques used, an operational RD optimization mechanism distributes the available bit rate among the techniques during encoding such that the predicted perceived distortion is minimal. This distribution is obtained by checking several bit-distribution options in an analysis-by-synthesis type procedure, and selecting the one with lowest distortion. The distribution can vary over time. Thus the bit allocation over the coding techniques is not fixed a priori, but flexible, allowing adaptation to the (time-varying) input signal and (time-varying) bit rate.

The coding techniques work in succession, that is, the first technique works on the original signal, the second one on the residue of the first one, and so on. The techniques can therefore encode parts of the signal that overlap both in time and in frequency. For a discussion of this multistage coding approach, see [25].

The coders work in a particular order for the complete duration of the signal. In our framework we made a preselection of coder order, because certain coder orders are more logical than others, for example, to use model-based (CELP, sinusoidal) coders to describe certain signal features first and then apply more generic waveform (transform) coders. In addition, complexity is reduced significantly because only a single coder order needs to be optimized instead of all possible coder orders. Note that the framework does not prohibit optimization of the coder order in any way, but this has not been investigated within this study for the reasons given.

The bit-rate allocation is done within an optimization frame the length of which can be varied. This allows the coding framework to adapt to delay constraints. Given that a constant bit budget is used for each coding frame, the length of the optimization frame is the fundamental cause for algorithmic delay (assuming that further processing such as the optimizations and sequential coding can be made in real time). The longer the delay, the more degrees of freedom the optimization algorithm has and the higher the coding efficiency. The segmentation of the coding techniques is aligned within the frames. Since such global segmentation is typically too large for the coding techniques, they also have their own segmentation, as visualized in Fig. 5 and described in the Section 1.3.

The bit-distribution algorithm spends the bits according to the strengths of the coding techniques. For example, in the hybrid configuration CELP and transform, the CELP coding technique plays the dominant role for speech signals, whereas for music signals the transform coding technique dominates. Nevertheless, for these "extreme" signals there is also an advantage of combining the two. As a result this hybrid configuration performs well for both speech and music signals. For more detailed results see [26]. Another hybrid example is the combination of sinusoidal and transform coding. In this case the sinusoidal

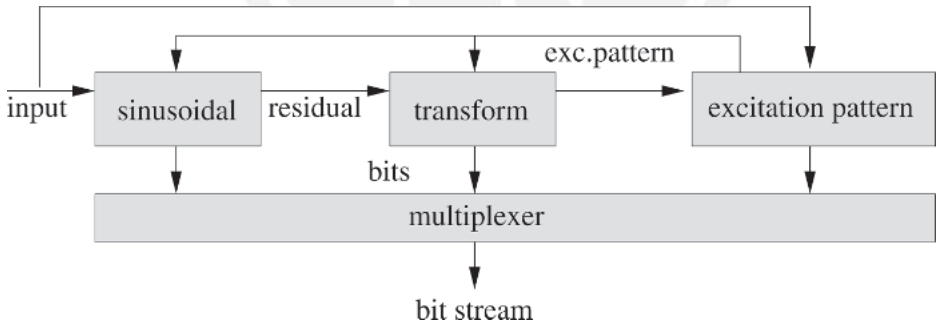



Fig. 4. Illustration showing position of excitation pattern coding.

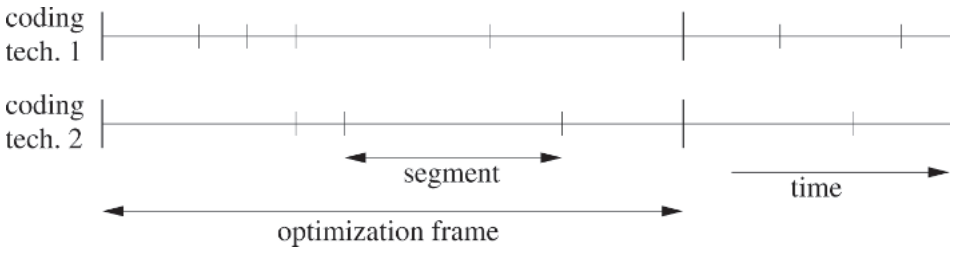



Fig. 5. Time aspects of hybrid coding framework.

coder codes the tonal components—its speciality—and then lets the transform coder handle the more spectrally flat residue, which is a better suited input signal for this technique [27], [25]. In this way the two techniques complement each other nicely, improving coding efficiency. Apart from adaptation to the input signal, this hybrid configuration can also handle the need to tailor the bit distribution to bit-rate constraints. For low bit rates our algorithm invests relatively more bits in the sinusoidal coder, which is model based and therefore more efficient at these bit rates than the transform coder, but at high bit rates the transform coder is outperforming the sinusoidal coder, and correspondingly relatively more bits are given to the transform coder. This behavior of the hybrid sinusoidal and transform coder is described in [28]. Figures about the bit allocation over the coding techniques will also be given in Section 2.

### 1.3 Segmentation and Distribution of Bit Rate over Segments

As described before, the optimization frame length determines a global segmentation, which is the same for all coding techniques. In addition, within this frame, each coding technique employs its own segmentation into variable-length segments (see Fig. 5) in order to adapt to local signal characteristics and bit-rate constraints.

In nearly all current coding techniques the input signal is segmented into time intervals that are coded separately. The optimal length of such segments depends on the available bit budget and on the input signal. For a long stationary signal a long segment is usually more efficient than a sequence of short segments, whereas for rapidly varying signals short segments are usually preferred. Note that coding efficiency is not only determined by the segment length, but also by its starting and ending point. Preferably these should be selected such that the signal is stationary within the segment. Specifically when frequency domain encoding techniques are used, such as transform coding or sinusoidal coding, stationarity within a segment will tend to lead to a more sparse spectrum, which can be encoded more efficiently. Due to these dependencies of coding efficiency on segmentation, the optimization algorithm can select a flexible segmentation that is often more efficient than uniform segmentation. For CELP coding on the other hand, which encodes signals in the time domain, less advantage is expected from flexible segmentation, since stationarity of signals is exploited using long-term predictors.

In this study segment length is defined as the distance between the crossover points of neighboring segments, also known as update rate. With many coding techniques overlapping windows are employed. So in the case of a 50% overlap, a segment length of 10 ms corresponds to a window length of 20 ms.

In AAC running at a sampling frequency of 48 kHz, for example, there are two options—a long segment of 21 ms or several short ones of 2.7 ms. A long segment is split into eight short ones in case this segment contains a transient. For GSM codecs, on the other hand, the segment length does not depend on the input signal, but different parameters have different update rates, such as 20 ms for the update of LPC coefficients and 5 ms for the gains.

The segmentation algorithms as employed within the framework under study are more flexible and based on those described in [29], [30], [17], [18], which use dynamic programming (Viterbi) [31], [32]. The generic algorithm allows for finding the segmentation $q$ of resolution $M$ that minimizes a cost function $J$. In our framework $J$ is a Lagrangian of perceptual distortion $D$ and bit rate $R$,

$$J(\lambda, q) = D(q) + \lambda R(q), \qquad \lambda \geq 0. \tag{1}$$

Given a target bit rate $R_t$, the optimal Lagrange multiplier $\lambda$ is found by maximizing $J(\lambda, q) - \lambda R_t$. Consider a signal that is initially divided into $N$ time units of $M$ samples. Let $J_{k,l}$ denote the cost for the interval $s_{k,l} = [kM, lM - 1]$, that is, the segment that consists of time units $k$ to $l$. Then at each iteration $i = 1, \ldots N$, the best segmentation of the interval $[0, iM - 1]$ is found by solving

$$J_i^* = \min_{0 \leq k < i} (J_k^* + J_{k,i}) \tag{2}$$

where $J_i^*$ is the minimum cost for the interval $[0, iM - 1]$. The minimizing argument of Eq. (2), say $k_i^*$, given by

$$k_i^* = \arg\min_{0 \leq k \leq i} (J_k^* + J_{k,i}) \tag{3}$$

is recorded as a split position and determines the optimal segmentation. The algorithm terminates once $J_N^*$ has been found, and the optimal segmentation can easily be determined by backtracking the optimal split positions. An example of this procedure is shown in Fig. 6 for $N = 3$.

Along with the segmentation, the allocation of bits over the segments is also determined by RD optimization principles, introducing an extra degree of freedom in the optimization formulas. Per optimization frame, a certain bit budget is allocated, and this budget is then distributed over the available coding techniques and segments. This makes it possible to spend most bits in segments that contribute most to the perceived quality instead of in segments that are less important, such as segments with (almost) silence. This is yet another possibility to adapt to the input signal, and the resulting variable bit rate leads to flexibility and increased coding efficiency. Moreover, the framework allows using a different bit budget for each optimization frame, which enables adaptation to changing channel capacities.

In the following more detailed information about the segmentation of the different coding techniques is given.

The sinusoidal coding technique uses a flexible segmentation similar to the one described. Finding the best segmentation is done efficiently by using dynamic programming techniques. The use of these techniques, however, implies the introduction of a long algorithmic delay since the optimal segmentation is only known after observing the complete excerpt, similar to the bit-distribution optimization. A solution to this problem has recently been presented in [33] and [34], where a suboptimal time segmentation was found within a specified time span. The loss in overall quality, however, turns out to be limited for

sinusoidal coding. Above 20 ms such a constraint does not have an important influence on the sound quality for sinusoidal coding [34]; effects are small between 20 and 50 ms, and above 50 ms they are negligible.

The transform coding technique also employs a dynamic-programming-based flexible segmentation algorithm, similar to the algorithms described in [29], [30], and [17]. However, since the transform windows typically have tail shapes that vary with the segment length, asymmetric transition windows are required at segment transitions if a nonuniform segmentation is desired. The use of such transition windows leads to a dependency between neighboring segments in the segmentation algorithms and thus requires customization of the existing segmentation algorithms. The approach taken in [35] aims at neglecting the dependencies that result from the use of transition windows; that is, the initial computation and optimization is done with regular symmetric transform windows. Once the segmentation has been obtained, an additional transform and coding step is performed for the given segmentation, where the appropriate transition windows are applied. This approach leads to a suboptimal segmentation. The actual performance loss incurred when neglecting the dependencies is analyzed in [36]. There it is shown that an optimal segmentation, in which all dependencies are taken into account, can still be found within polynomial time and that a slight increase in performance can be observed. Nevertheless the complexity increase may still be significant, and it is justified to neglect the dependencies that result from using transition windows.

Apart from flexible time segmentation, the transform algorithm can also use two flexible frequency decomposition algorithms. In [37] the use of signal transforms that lead to a nonuniform frequency decomposition is discussed. These transforms are obtained by employing subband merging [38], a technique originally devised to construct nonuniform cosine-modulated filterbanks. Subband merging allows for a time–frequency tradeoff within a single transform window. Using the aforementioned dynamic-programming algorithms, a flexible frequency decomposition into nonuniform subbands is obtained in [37], where an operational RD control selects the optimal frequency decomposition and the corresponding bit allocation. While this method for creating nonuniform frequency decompositions may provide valuable insights into the signal characteristics, the method suffers from a high side-information rate for coding the decompositions, which reduces the efficiency of the algorithm. The work in [39] incorporates a second technique that leads to nonuniform frequency decompositions, that is, frequency-domain linear prediction or temporal noise shaping [40]. Similar to the work in [19], operational RD is used to select the prediction filter orders and bit allocation that lead to the lowest perceptual distortion for a given target bit rate. Through listening tests it is shown that this algorithm can increase the coding efficiency substantially [39].

Whereas the other coding techniques basically encode structures in the frequency domain (and segmentation helps to create spectral structures that can be encoded efficiently, as also mentioned in Section 1.3), the CELP coding technique encodes structures in the time domain. Therefore flexible segmentation has less of an advantage for this technique and was not used. However, the CELP coding technique did have a flexible allocation of bits over the segments (variable bit rate; see Section 1.5.3).

### 1.4 Perceptual Distortion Measures

The perceptual distortion measures are important components of the operational RD optimization mechanisms because they provide predictions of the perceptual distortion or, equivalently, sound quality as perceived by a listener. With this information the optimization mechanisms aim to code the perceptually relevant signal features as well as possible, not wasting bits on irrelevant signal components. To predict the perceived distortion, the distortion measures need as inputs the original and the synthesized signal resulting from particular coding settings. The output is a scalar that quantifies the distortion. The optimization mechanism will of course choose those settings that result in the lowest distortion for the available bit budget. As a consequence the accuracy of the perceptual distortion measure is of vital importance, determining the quality of the encoding.

To determine the optimal segmentation and bit allocation per segment, a relatively simple distortion measure is used, which utilizes the spectral auditory masking properties of the input signal. This spectral distortion measure uses a frequency-domain filterbank that mimics peripheral

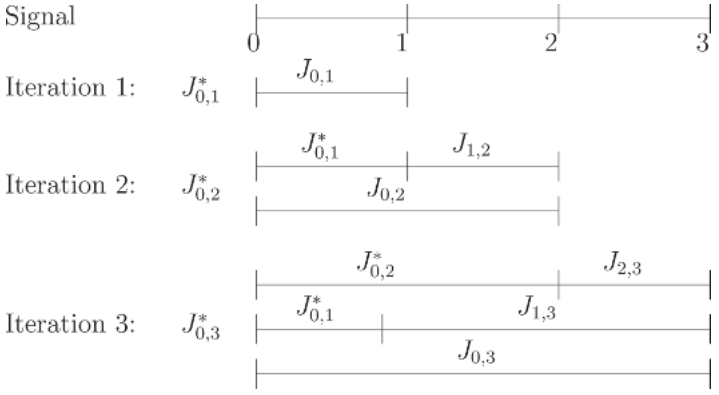



Fig. 6. Flexible time segmentation algorithm using dynamic programming to build up optimal segmentations.

auditory filtering. Within each filter the excitation is determined resulting from the original (masking) signal, creating an excitation pattern across all filters, which, for the simple distortion measure, captures all information necessary to determine the spectral masking properties of the original signal. When a certain distortion signal is introduced due to the coding operation, the error signal power is calculated within each auditory filter and divided by the excitation power due to the original signal. This ratio provides an estimate of the perceived distortion within each auditory filter. By assuming that perceived distortions are integrated across auditory filters (see [41]), we obtain an estimate of the overall perceived distortion (see [42] for more detail).

To determine the optimal distribution of bits over the coding techniques, a more complex distortion measure needs to be used that takes into account the temporal masking characteristics of the auditory system in addition to the spectral masking characteristics. This measure will be called the spectrotemporal distortion measure and is based on the auditory model of Dau et al. [43]. Because this model was developed to simulate listener behavior in psychoacoustical experiments, computational complexity was not an important issue. When applied in the context of coding, where for each segment the perceptual distortion has to be determined many times, complexity becomes an issue. For this reason the spectrotemporal model was changed into a two-stage model, where first the masking properties of the original signal are determined once, followed by a low-complexity stage, which determines the perceptual distortion for each coding decision that is considered, reusing the masking properties of the original signal.

## 1.5 Coding Techniques

### *1.5.1 Sinusoidal Coding*

One coding technique used in the framework is sinusoidal (parametric) coding, which is model based and therefore particularly efficient for audio at low bit rates.

*Parameter Estimation* For the extraction of sinusoids, a psychoacoustic matching pursuit [20], [44], [45] is used. This algorithm nicely fits into the framework's philosophy of operational RD optimization by selecting, in each operation, the sinusoidal component that is perceptually most relevant (for a review of this method and related ones, see [46]).

With a basic sinusoidal coder, that is, one using constant-amplitude basis functions, sudden changes (transients) in the signal cannot be modeled very efficiently. This can be understood intuitively by looking at the underlying model, which consists of (stationary) sinusoids. By applying adaptive segmentation using RD optimization, transients can still be handled, but to improve the modeling of transient signals further, the algorithm can be extended with amplitude modulation (AM). A number of adapted signal models based on AM have been proposed specifically for dealing with transients, and a number of coders have been developed using these models [47]–[51] (see also [52]). The models allow the individual sinusoidal components to have time-varying envelopes within segments. The papers referenced differ in the model they impose on this envelope. Since these models have additional parameters associated with them and since most audio segments are stationary, these modified models will not always be the best choice. However, any heuristic switching is avoided by RD optimization. In [50] an amplitude-modulated sinusoidal audio coder was presented that was based on a nonlinear model of the modulating signal, characterized by an onset time, an attack parameter, and a decay parameter (for definitions see [50]). This model was demonstrated to improve greatly on the perceived quality in a MUSHRA-like test (for an explanation of the MUSHRA test, see Section 2.2.1) using critical transient excerpts [50]. Specifically, the listening test showed average improvements of 10 points on a MUSHRA scale compared to the coder with no AM and more than 30 points for some individual signals. The results reported in [50] prove that it is indeed efficient in terms of bit rate to allow different modulating signals for different components and that optimal segmentation and adapted models are complementary coding techniques. Furthermore, the optimal segmentation changes with the signal model. In addition two other AM coders have been developed. In [49] the amplitude modulating signal is modeled as a linear combination of arbitrary basis vectors. This coder was demonstrated in a preference listening test to improve upon a sinusoidal coder, but although highly flexible, it suffered from a high complexity. An implication of the AM coders and their complicated signal models is the computational complexity associated with finding the parameters. An amplitude-modulated sinusoidal audio coder based on the theory of [47] and the results of [48] was developed in [51]. It uses frequency-domain linear prediction as a means of estimating and efficient coding of the envelopes in critically sampled subbands. This coder has very low complexity and requires little memory compared to that of [50], and it was demonstrated in a MUHSRA-like listening test to improve upon a baseline coder, that is, one using constant-amplitude sinusoids [51].

*Parameter Quantization* In order to represent sinusoidal components in the bit stream efficiently, sinusoidal parameters need to be quantized. This quantization is also based on RD principles. The scalar quantizers for sinusoidal amplitudes, frequencies, and phases are optimized jointly, such as to minimize, given a bit-rate constraint, a perceptual distortion measure. As a result each and every sinusoidal parameter is quantized differently, more accurately for perceptually more relevant components at the expense of less relevant ones. The perceptual importance of components is defined by the masking curve, which is available at both the encoder side and the decoder side as a result of the transmission of the excitation pattern (see Section 1.5.4).

The jointly optimal quantizers for sinusoidal parameters are found analytically using high-resolution, or high-rate, quantization theory. The analytical solutions make it possible to design quantizers at very low computational complexity. High-resolution theory assumes that probability

density functions (pdf) of input parameters can be approximated accurately as being constant within quantization cells. Quantizers are expressed in terms of quantization point densities, which for scalar quantizers are the inverses of the quantization step sizes. The application of high-resolution quantization theory to sinusoidal coding was first presented for amplitudes and phases in [53], [54] and further extended to frequencies in [55]–[58].

In this work the quantization of sinusoidal parameters is implemented according to quantization method 3 of [56]. We minimize the following weighted mean-squared error (WMSE):

$$D = \frac{1}{L}\sum_{l=1}^{L} \upsilon_l D_l \tag{4}$$

where $L$ denotes the total number of sinusoids, $\upsilon_l$ is the perceptual weight defined as the reciprocal of the masking curve evaluated at frequency $\omega_l$, and $D_l$ is the per-sample MSE evaluated over a segment of length $N$ samples. Using the high-resolution assumption, the per-sample MSE can be written as [54], [56]

$$D_l \approx \iiint f_{A,\Omega,\Phi}(a, \omega, \phi) \times \left[\frac{g_A^{-2}(a)}{24} + a^2\left(\frac{g_\Omega^{-2}(\omega, a)N^2}{288} + \frac{g_\Phi^{-2}(\phi, a)}{24}\right)\right] da\, d\omega\, d\phi \tag{5}$$

where $f_{A,\Omega,\Phi}$ $(a,\ \omega,\ \phi)$ is the joint pdf of amplitude $a$, frequency $\omega$, and phase $\phi$ and $g_A$ $(a)$, $g_\Omega(\omega, a)$, and $g_\Phi$ $(\phi, a)$ are the quantization point densities of amplitude, frequency, and phase, respectively. The distortion [Eq. (4)] is minimized subject to an entropy constraint

$$H = \frac{1}{L}\sum_{l=1}^{L} H_l \tag{6}$$

where $H_l$ is the joint entropy of amplitude, frequency, and phase quantization indices of the $l$th sinusoid. Using the high-resolution assumption, the joint entropy of the quantization indices is approximated as [54], [56]

$$H_l \approx h(A, \Omega, \Phi) + \iiint f_{A,\Omega,\Phi}(a, \omega, \phi)\{\log_2[g_A(a)] + \log_2[g_\Omega(\omega, a)] + \log_2[g_\Phi(\phi, a)]\}\, da\, d\omega\, d\phi \tag{7}$$

where $h(A,\ \Omega,\ \Phi)$ is the joint differential entropy of amplitude, frequency, and phase.

We quantize sinusoidal amplitudes relative to the masking curve, that is, quantizers are designed for normalized amplitudes $\tilde{a}_l = a_l \upsilon_l^{1/2}$. The optimal quantization point densities are then given by

$$g_\Omega(\omega) = g_\Omega = \left[\frac{N^2\sigma(\tilde{A})}{12}\, 2^{\tilde{H}-b(\tilde{A})}\right]^{1/3} \tag{8}$$

$$g_{\tilde{A}}(\tilde{a}) = g_{\tilde{A}} = \left\{\left[\frac{N^2\sigma(\tilde{A})}{12}\right]^{-1/2} 2^{\tilde{H}-b(\tilde{A})}\right\}^{1/3} \tag{9}$$

$$g_\Phi(\phi, \tilde{a}) = g_\Phi(\tilde{a}) = \tilde{a}\left\{\left[\frac{N^2\sigma(\tilde{A})}{12}\right]^{-1/2} 2^{\tilde{H}-b(\tilde{A})}\right\}^{1/3} \tag{10}$$

where $\tilde{H} = H - h(\tilde{A}, \Omega, \Phi)$, $b(\tilde{A}) = \int f_{\tilde{A}}(\tilde{a}) \log_2(\tilde{a})\, d\tilde{a}$, and $\sigma(\tilde{A}) = \int f_{\tilde{A}}(\tilde{a})\, \tilde{a}^2 d\tilde{a}$. The optimal quantizers are uniform and, therefore, easy to implement. The parameters $h(\tilde{A}, \Omega, \Phi)$, $b(\tilde{A})$, and $\sigma(\tilde{A})$ are estimated a priori and stored at the encoder and the decoder.

High-resolution theory can facilitate a number of further optimizations and flexible analytical solutions in coding applications. In [59], [60] it is used to facilitate the joint optimization of variable-length time segmentation, distribution of sinusoidal components over segments, and quantization of sinusoidal parameters. In [61] differential quantization of sinusoidal parameters is presented. In [62] it has been shown that the quantizers can be improved by the joint quantization of multiple sinusoids as opposed to quantizing each sinusoid individually. This work has also been extended to multiple descriptions for packet-based networks [63]. In [27], [25] high-resolution theory is used to find jointly optimal quantizers in multistage audio coding, where a sinusoidal coder is combined with a transform coder, such as a modified discrete cosine transform coder.

### 1.5.2 Transform Coding

Another coding technique of the framework is transform coding. Since this technique is waveform based, it has strengths that are basically complementary to those of sinusoidal coding, transform coding being most efficient for sound at high bit rates. The transform coding technique uses the modified discrete cosine transform (MDCT) because this transform has many desirable properties for sound coding such as critical sampling, minimal blocking artifacts, good channel separation, and perfect reconstruction in the absence of quantization. Different transform lengths are used within the transform coder: 256, 1024, 2048, and 4096 samples. The short window length (256 samples) is effective for transient signals, whereas the longer windows are more likely to be used for stationary signals. The very long window with 4096 samples is especially effective for more stationary signals and a very low target bit rate. The appropriate window lengths within an optimization frame of the signal are determined by the RD optimization (see in the preceding).

The resulting transform coefficients are quantized according to RD principles. In other words, being beneficial from a perceptual point of view, coarser quantization is used for the coefficients where the masking curve indicates a higher masking effect—quantization is less critical—and vice versa. This is obtained by weighting each coefficient with a scale factor $\Delta[m]$, which is obtained from the masking threshold $M[\cdot]$ at the corresponding frequency $m$ and one global scale factor $g_{\mathrm{scf}}$ that applies to all frequencies,

$$\Delta[m] = g_{\mathrm{scf}} \cdot \sqrt{12 \cdot M[m]}. \tag{11}$$

Afterward all weighted coefficients are quantized using the same quantizer, a larger scale factor leading to coarser quantization. This global scale factor is therefore used to control the overall quality and bit rate for the segment and is obtained/controlled by the RD optimization. In other

words, by adaptation of the global scale factor for each segment the target rate is reached.

A second function of the global scale factor is to determine the cutoff frequency of a low-pass filter, which is applied before quantization [64]. Higher global scale factors lead to lower cutoff frequencies. This is to control the tradeoff between audible coding artifacts due to a coarse quantization and the coded bandwidth of the signal at a given bit rate. This control is needed because distortions that are related to modulation of the bandwidth across segments are not captured by the applied distortion measure, because it operates on a single-segment basis.

For quantization a four-dimensional regular lattice vector quantizer with nonuniform quantization is used [65]. Afterward the quantized vectors are Huffman coded using a dedicated technique, which is described in [64]. This coding scheme is particularly effective for low bit rates, at which many zero coefficients have to be coded. A more detailed description of the transform coding technique can be found in [64].

Based on the preceding, a separate coding technique for transients has been developed. This multistage technique is based on the detection and extraction of transient components in a short-window MDCT spectrum. The quantization of these components is done according to the quantization scheme described, that is, using a global scale factor determined by RD optimization and considering masking effects. In addition the sparseness of the transient spectra is taken into account. The idea behind this transient coder is that the residue, which results after subtracting the transients and which is therefore free of transients, can afterward be coded with an optimized coding technique for stationary signals, such as transform coding using only long windows. More information about this method can be found in [66].

### 1.5.3 CELP Coding

The third coding technique that has been used is CELP coding, which is based on predictive coding. Again, the strengths of this technique are complementary to those of the other techniques, CELP coding having exceptional efficiency for speech at low bit rates by its use of a speech model. Our CELP coding technique has been derived from the lower band (50–6400-Hz) part of the standard GSM/3GPP AMR-WB codec [67]. The segment length of 256 samples has been kept, but the module works at 12.0-kHz sampling frequency instead of 12.8 kHz.

The CELP module has been extended with an operational RD mechanism. For each segment nine RD operating points are obtained ranging from 0 to 21.75 kbit/s (including extra bits needed for rate signalization) using the spectral distortion measure as described in Section 1.4. Thus a variable bit rate is obtained with a corresponding improved coding efficiency.

Adaptations have been made such that the CELP coding technique can function together with the other coding techniques, because the CELP method needs to maintain state variables (past input and output signals). These are not always available, for example, in case the CELP codec was not used in the previous optimization frame, and then special provisions are taken [26].

The CELP codec encodes a limited bandwidth (6-kHz) signal while the missing higher band can still contain important information. A bandwidth extension module has therefore been designed for the CELP codec that is activated at the decoder when the CELP codec is the first coder in the cascade. This feature operates in a similar way as that in the AMR-WB+ codec [8]. The integration of the CELP coder in the framework is detailed in [26].

### 1.5.4 Excitation Pattern Coding

Apart from the parameters of the coding techniques, additional information is included in the bit stream—the excitation pattern. This is a perceptually relevant representation of the spectral envelope of the original signal and is a model for the spectral energy distribution across auditory filters in the human auditory system [24]. The pattern is calculated for short segments of the input signal so that it can quickly adapt to changes over time. The excitation pattern information has two important functions. First it enables the derivation of a masking curve at the decoder side. Second it can be used to complement the signal part that is generated by the coding techniques with a synthetic noise signal.

The masking curve is derived from the excitation pattern by determining for each frequency the threshold level a probe tone should have in order to be just detectable, given the excitation pattern resulting the original signal. Since the excitation pattern information enables the generation of a masking curve both at encoding and at decoding, the coding techniques have the opportunity to code their parameters relative to this masking curve without coding this curve themselves. This increases the coding efficiency and is therefore exploited by the sinusoidal and the transform coding techniques.

Furthermore the excitation pattern is used to complement the signal part of the coding techniques with a synthetic noise signal to account for the missing parts in the spectrum. In many situations the available bit rate is too low to model the entire spectrum, which results in lower signal energy in certain frequency regions and hence a decrease of part of the excitation pattern. These missing parts are then modeled with spectrally shaped noise, which leads to better quality compared to the alternative of not modeling these parts of the spectrum at all. To determine the noise part, both the excitation pattern of the original signal, that is, the excitation pattern information in the bit stream, and the excitation pattern of the signal synthesized by the coders, which is calculated in the decoder, are needed. Then a noise signal is generated such that the total of synthesis signals of the decoders plus the noise signal result in an excitation pattern equal to the original excitation pattern. This process is described in more detail in [68].

To allow the transmission of the excitation patterns in the bit stream a compact representation of these parameters is necessary. For example, differential pulse code modulation (DPCM) coding of the properly quantized ex-

citation patterns calculated every 128 samples (at a sampling frequency of 48 kHz this corresponds to a new excitation pattern every 2.67 ms) consumes as many as 45 kbit/s. This is far too much, compared to the bit rate needed for the scale factors of, for example, an AAC audio coder, which is approximately 4 kbit/s out of a total bit rate of 48 kbit/s [69]. Therefore an optimized coding algorithm for the excitation patterns was developed.

A number of subsequent excitation patterns is taken together in an excitation pattern matrix. This matrix is transformed using a two-dimensional discrete cosine transform (DCT). Next the transform coefficients are quantized and entropy coded. As the relevant transform coefficients are not localized in a fixed region, due to differences between transient and stationary signals, a flexible algorithm is needed, which automatically covers all these relevant coefficients and applies a suitable quantization and coding. This is efficiently achieved by the set partitioning embedded block (SPECK) coder, which was effectively applied in image coding [70].

To further improve the coding efficiency for the excitation patterns, additional linear prediction from the preceding excitation pattern matrix to the actual one is performed. With this coding scheme it is possible to code the excitation patterns at a suitable accuracy with a bit rate of approximately 4 kbit/s. Therefore the coded excitation patterns can be used as side information for low-bit-rate coding. A more detailed description of the coding algorithm of the excitation patterns can be found in [71] and [64].

#### *1.5.5 Lossless Coding*

After parameterization of the input signal and quantization of the resulting parameters, another coding step is performed, lossless encoding, which removes the redundancy from the quantized parameters. For this lossless compression, Huffman coding [72] is used with one fixed set of Huffman tables. So for all input signals, bit rates, and so on, the same set of parameter statistics has been used. These tables have been generated beforehand based on the parameter statistics across many sound signals. If during encoding a certain entry in the precalculated Huffman code word table appears to be missing, an escape code is used, which happens only rarely.

In order to reduce the bit rate needed for the sinusoidal parameters, the inter- or intrasegment correlation between sinusoidal parameters can be exploited using differential or predictive coding schemes. For example, [3], [4], [73], [74] apply time-differential (TD) techniques, where sinusoidal components in the current segment are represented using parameter differences or ratios relative to components in the previous segment. By doing so, the probability density function of the relative parameters is much more peaked which results in a significant coding gain. Similarly, [75] applies frequency-differential (FD) techniques to represent sinusoids in a given segment relative to the sinusoids in the same segment. In [76] it was shown that when combined with variable segmentation, sinusoidal FD coding can be as RD efficient as TD encoding. FD techniques, however, have the additional advantage that they are robust to packet losses when used in packet-based transmission channels. Optimal entropy-constrained quantization schemes for differential parameters have been presented in [61], which are valid for both the inter- and intrasegment situation.

### 1.6 Stereo Coding

The RD optimized coding framework described in the preceding section is mono, and since for many audio coding applications stereo or even multichannel coding is required, it has been investigated how this framework can be extended to more channels. This can be obtained by adding stereo preprocessing and postprocessing (see Fig. 7) Being independent of the mono coding, the stereo coding extension follows a hybrid approach combining parametric stereo coding with waveform coding, enabling a seamless transition from low to high bit rates and convergence to transparent quality.

The stereo coding framework is based on sum–difference coding of signals that are aligned in time. After time and gain aligning the right and left channels, two signals are formed—a sum signal, which is coded with the mono coder, and a residual (difference) signal. The time and gain alignment parameters capture the coarse stereo image and may be transmitted using only a few kilobits per second, implementing a pure parametric stereo extension. By encoding the residual signal using a waveform encoder and an increasing number of bits, the stereo image is encoded with finer and finer detail, and convergence toward transparent quality is obtained.

The delay-compensated sum–difference stereo coding framework is used in a perceptual filterbank structure and generalizes to the multichannel (more than two channels) case in a straightforward way. See [77] for more information about this stereo coding framework and how it relates to other methods, such as mid-side coding [78] and parametric stereo coding [79]–[81]. Incorporation of the stereo coding part into the RD optimization framework is still subject to future research.

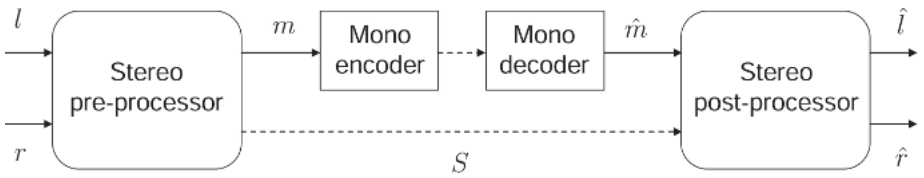



Fig. 7. Stereo coding using pre- and postprocessing techniques. —— signals; - - - parameter bit streams. *l:* left channel; *r:* right channel; *m:* time- and gain-matched sum of *l* and *r;* $\hat{m}$*:* decoded sum of *l* and *r;* $\hat{l}$*:* decoded left channel; $\hat{r}$*:* decodecd right channel; *S:* stereo extension bit stream.

### 1.7 Complexity

Because of the operational RD optimizations, the encoding process of our framework is inherently complex. This applies in particular to those optimizations that use a closed-loop approach, that is, for each iteration, the encoding, decoding (synthesis), and distortion calculation are done to deliver the required RD points for the optimization. This is done, for example, to decide on the bit distribution among the coding techniques. The complexity is not so much in the search among candidate settings, but in the initialization thereof, namely, the generation of the RD points among which the search has to be performed. This initialization, however, can be parallelized. Furthermore, there are several other ways to lower complexity; here we describe three.

First, instead of the closed-loop approach, an open-loop approach can be used [82]–[84]. This approach does not require encoding, decoding, and distortion calculation for each iteration, but instead predicts the RD points, reducing complexity considerably. Fig. 8 illustrates this. First the input signal is reduced to a property vector: a high-level descriptor of the properties of the signal containing information, such as loudness and spectral flatness, that is sufficient for a required predictor accuracy. From this property vector the distortion resulting from a particular coding configuration, corresponding to a certain rate, is predicted.

The selection of a property vector from the input segment is of great importance for the performance of the proposed framework. The selected property vector should be a representative for the incurred distortion in the current segment for the given coder. In more theoretical terms, the random input segment $\boldsymbol{s}$ is processed by the encoder into the distortion variable $D$ and by the property extractor into the property vector $\boldsymbol{P}$. The basic task for the property extractor $f(\cdot)$ is to extract properties $\boldsymbol{P}$ that contain sufficient information about $D$ for a required predictor accuracy. The amount of information that $\boldsymbol{P}$ contains about $D$, or the suitability of a given property vector, can be measured by the mutual information $I(D; \boldsymbol{P})$ as discussed in [82]. A more practical approach to the selection of a good property vector, based on a "deflation" strategy, was proposed in [84]. Similar to pruning, the idea is to start out with a large number of properties and then iteratively remove components until the estimation performance starts to degrade over a given test set.

The second part of the open-loop approach is the predictor. The aim of the predictor $g(\cdot)$ is to find a prediction $\hat{\delta}$ of the incurred distortion $\delta$, based on an observation of the property vector $\boldsymbol{P} = \boldsymbol{p}$. In [82] a model-based prediction approach, based on Gaussian mixture models (GMM), was proposed. Utilizing a pretrained GMM for the joint distortion property pdf $f^{(M)}_{D,\boldsymbol{P}}(\delta, p)$, we approximate the minimum mean-squared error (MMSE) at each coding instant as

$$\hat{\delta} = g(\boldsymbol{p}) = \int \delta f^{(M)}_{D|\boldsymbol{P}}(\delta \,|\, \boldsymbol{P} = \boldsymbol{p})\, \mathrm{d}\delta \tag{12}$$

where $f^{(M)}_{D|\boldsymbol{P}}(\delta|\boldsymbol{P} = \boldsymbol{p})$ is the conditional model pdf, which can be shown to be a mixture of Gaussian densities and is easily derived from the joint model pdf $f^{(M)}_{D,\boldsymbol{P}}(\delta, \boldsymbol{p})$. In practice this predictor calculates a weighted sum of conditional means,

$$\hat{\delta} = \sum_{i=1}^{M} \rho_i' \boldsymbol{m}_{i,D|\boldsymbol{P}=\boldsymbol{p}} \tag{13}$$

where $M$ is the number of mixture components, and $\{\rho'_i\}$ and $\{\boldsymbol{m}_{i,D|\boldsymbol{P}=\boldsymbol{p}}\}$ represent the weights and the means of the conditional model pdf $f^{(M)}_{D|\boldsymbol{P}}(\delta|\boldsymbol{P} = \boldsymbol{p})$, respectively.

This distortion prediction approach has been tested for various coders and coding scenarios [82]–[84], and it was reported in [84] that this approach can lead to a complexity reduction by a factor of 10 compared to a closed-loop approach.

Second the complexity of the spectrotemporal distortion measure can be reduced by means of a sensitivity matrix. Under the assumption of small errors the distortion measure $d(x, \hat{x})$ between the original signal block $x$ and a coded candidate $\hat{x}$ can be approximated by a weighted quadratic norm on the error signal [85],

$$d(x, \hat{x}) \approx \frac{1}{2}(x - \hat{x})^{\mathrm{T}} M(x)(x - \hat{x}). \tag{14}$$

The so-called sensitivity matrix $M(x)$ is obtained from a linearization of the psychoacoustical model for the current signal block [86], [87].

The reduction in complexity stems from the fact that the psychoacoustical model is only used once per block to calculate the sensitivity matrix $M(x)$ from the original signal block $x$, instead of having to be called repeatedly for every possible coded candidate $\hat{x}$. Furthermore matrix analysis on $M(x)$ can be used to obtain important information about the current state of the spectrotemporal model, such as a masking curve [86], [87]. This information can then be used in the different coding techniques.

Third the operational RD optimizations for the segmentation can be avoided by applying upfront segmentation of the signal. In this case the RD cost function [see Eq. (1)] is replaced by a cost measure that is independent of distortion and rate, such as perceptual entropy [88]. Once the segmentation has been obtained, the bit allocation for the presegmented signal can still be performed using operational RD optimization. While this separation of segmentation and bit allocation into two separate stages is suboptimal, [89] has shown both numerically and through listening tests that the loss of coding efficiency is negligible. The decrease in complexity, however, can be as high as 50%. While this method of upfront segmentation has only been studied for the transform coder in [89], it may be applied equally well to the sinusoidal coder. As such, for

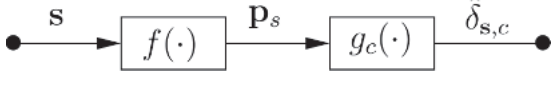



Fig. 8. Open-loop approach for operational RD optimization by predicting distortion for a particular signal $\boldsymbol{s}$ and coding configuration $c$. Property extraction function $f(\cdot)$ processes $\mathbf{s}$ into a property vector $\mathbf{p}_s$. Based on $\mathbf{p}_s$ and given $c$, prediction function $g_c(\cdot)$ predicts distortion $\hat{\delta}_{\mathbf{s},c}$.

hybrid coding situations, the resulting decrease in complexity can be substantial.

In summary, although the operational RD optimization approach results in high complexity of the encoder, there are several ways to reduce this, which is important for practical encoding applications. The operational RD approach does not increase the complexity of the decoder.

## 2 VALIDATION

### 2.1 Application Scenarios and Validation Design

The operational RD optimized hybrid coding framework has been validated in extensive listening tests. Three application scenarios have been considered for this validation: 1) broadcasting with (dynamic) adaptation to network resources, 2) flexible storage, and 3) enhanced communication.

The broadcasting scenario deals with streaming and transmission applications at very low bit rates and with situations where the bandwidth (bit rate) of the transmission network varies over time. Therefore the low-bit-rate performance of the adaptive RD optimized hybrid codec was evaluated at a bit rate of 20 kbit/s with the MUSHRA method (see Section 2.2.1). In order to measure the performance for varying bit rates, an ECQ evaluation test (see Section 2.2.2) was performed with various temporally varying bit-rate patterns where listeners gave continuous (instantaneous) ratings of sound quality.

The flexible storage scenario deals with storage applications that essentially only impose requirements with respect to the total amount of information that is associated with encoding a particular file. In a first test the performance of an adaptive RD optimized hybrid codec was investigated for a high-storage-capacity scenario using a bit rate of 64 kbit/s. Since quality was close to transparency, the ITU-R BS.1116 test was used (see Section 2.2.3). For low-storage-capacity scenarios a MUSHRA test was employed using the same codec at 24 and 48 kbit/s.

For the enhanced communication scenario an important constraint is the algorithmic delay. Since the RD optimization requires a certain optimization frame for determining the encoder settings, lowering the algorithmic delay can influence encoding efficiency. Two RD optimized hybrid coders were evaluated at two bit rates (20 and 48 kbit/s) using a MUSHRA test. The encoders were set at different algorithmic delays (43 and 384 ms) to compare performance.

In all evaluation tests relevant state-of-the-art encoders were added to serve as a reference to compare performance. In the broadcasting and communication scenario tests, both speech and music excerpts were employed, the storage scenario test used predominantly music excerpts. More information about the coder settings is given in Section 2.3.

### 2.2 Methods and Excerpts

#### *2.2.1 MUSHRA Method*

The MUSHRA (multi stimuli with hidden reference and anchor points) method [90] is dedicated to the assessment of intermediate quality. It has been recommended by the ITU-R as BS.1534. An important feature of this method is the inclusion of a hidden reference and bandwidth-limited anchor signals. In our test the chosen anchor points were the low-pass-filtered originals with cutoff frequencies of 3.5 kHz (mandatory) and 7 kHz. Each listener (expert in the audio domain) had a training session of about 15 min in order to get familiar with the test methodology and software and with the kind of quality they had to assess. This was also an opportunity to adjust the playback level that then remained constant during the test phase. As a randomization process was used, the order of the excerpts was different for each listener. Test instructions explained to the listeners how the software worked (the CRC-SEAQ software was used), what they were going to listen to (briefly), how to use the quality scale (1–100, bad to excellent), and how to score the different excerpts. It was also mentioned that there was a hidden reference signal to score, which was later used in the rejection process of listeners to verify whether this score was at least 90. The tests were performed on the headphone STAX Signature SR-404 (open model) and its amplifier SRM-006t. The digital sound was played through the PC board Digigram VX 222 and converted by a 24-bit DAC (3Dlab DAC 2000).

#### *2.2.2 ECQ Method*

The ECQ (continuous quality evaluation) methodology [91] has been standardized in ITU-T Q12/7 (recommendation P.880, 2004 May). It can be used for evaluating the impact of time fluctuations in artifact levels on the instantaneous perceived quality (that is perceived at any instant of a sequence) and on the overall perceived quality. The method uses a two-part task—first instantaneous (continuous) judgments during the sequence, and second an overall judgment at the end of the sequence.

In contrast to the P.880 recommendation to have at least 24 naive listeners participating in a test, we used 16 trained listeners for practical feasibility and for making the test as critical as possible. Prior to the test, subjects were trained by listening to two sequences. These sequences, 45 s long, were extracted from the items used in the test and covered different quality levels and different quality fluctuations representative of the range of temporal fluctuations and quality levels that the subjects encountered during the actual test.

For the continuous judgment an electronic slider connected to a computer was used for recording the instantaneous quality assessment from the subjects.[2] The initial slider position was always at about the midpoint of the scale. Subjects were instructed to assess the sound quality of the sequence continuously by moving the slider along the scale such that its position reflected their opinion on

[2]The slider device had the following characteristics: slider mechanism without any reset position (that is, no automatic return to a predefined position), linear range of travel of 110 mm, fixed on test desk. The slider position was recorded twice a second, and was coded from 0 (bottom of scale) to 255 (top of scale).

quality at that instant. Five labels were shown along the scale: Excellent, Good, Fair, Poor and Bad to help the subject associate the slider position with suitable ranges of sound quality.

For the overall judgment a set of five buttons, numbered from 1 to 5, was used. At the end of each sequence, subjects were asked to rate its overall quality on the five-category listening-quality scale also used in the continuous judgment.

#### 2.2.3 ITU-R BS.1116 Method

The ITU-R BS.1116 method [92] has been selected to test whether near-transparent quality can be achieved. In this method the listener compares two signals (a hidden reference and the signal under test) to the original signal and must decide which one is the signal under test. Then the listener has to score along a five-grade scale the way he perceives the degradation in this signal (5—imperceptible, 4—perceptible but not annoying, 3—slightly annoying, 2—annoying, 1—very annoying).

As for the MUSHRA method, there was a training session of about 15 min in order to familiarize the listeners (experts in the audio domain) with the test methodology and software and with the kind of quality they had to assess. Listeners also had the opportunity to adjust the playback level. Sound was played on Beyerdynamic DT 990 Pro headphones using a Marantz CDA-94 digital-to-analog converter.

#### 2.2.4 Test Items

The sampling frequency was 48 kHz for all excerpts. The duration of the items ranged from 7 to 15 s. All excerpts were normalized in amplitude at 80% of the full digital scale in order to avoid too large differences in loudness. Fade in and fade out were applied if necessary.

For the broadcasting test dedicated to streaming and transmission, the test set contained more speech than music. The items selected for this test were chosen to be realistic types of excerpts as much as possible, keeping in mind that they should remain as critical as possible as well (that means that transparency is often not achieved by state-of-the-art encoders when encoding those sequences). In order to limit the duration of the test, only five items were chosen. These are described in Table 1. For the ECQ methodology only two items were chosen: one more speech oriented (commentary of a basket ball match with applause and people shouting, 70 s) and another containing music (jazz music with a female singer in English, 90 s).

For the storage test dedicated to high capacity, the test set contained more music than pure speech. This made this test more music oriented. Again the items were chosen to be realistic types of excerpts as much as possible, keeping in mind that they should remain as critical as possible as well. In order to lighten the test, only six items were chosen, as listed in Table 2.

For the enhanced communication test the test set contained more speech than music, which made this test more speech oriented. The first four excerpts from Table 1 were used.

### 2.3 Coder Configuration

The encoder configuration that was used in the evaluation tests was kept constant as much as possible, varying only the constraints (bit rate and algorithmic delay) imposed by the application scenario. The RD optimization was left to determine many of the detailed encoder decisions according to its optimality criteria. One aspect of the encoder configuration was decided beforehand, depending on the bit rate, to restrict computational complexity. For bit rates below 24 kbit/s a cascade of CELP and transform coder was used because a CELP coder is highly efficient at low bit rates for speech signals. The transform coder was used as a residual coder and fallback option in case the CELP coder was inefficient (such as for music). For 24 kbit/s and higher bit rates, a cascade of a sinusoidal and a transform codec was used, where the sinusoidal codec was specially suitable for highly tonal music excerpts while the transform coder was used as residual coder for encoding the less tonal sound components. Furthermore the bit-distribution step was set to 25%, again for reasons of complexity. In other words, the bit-distribution options that were checked by the optimization mechanism were 100% to the first coder and 0% to the second coder, 75% to the first coder and 25% to the other codecs, and so on.

Optimization frames were 384 ms long in most cases, except for the communication scenario, where the frame size was 43 ms. These frames were used to optimize the bit-rate distribution between both codecs.

Within each frame the individual codecs could determine the optimal segmentation independently. For the sinusoidal codec, 11-, 16-, 21-, and 27-ms segments could be used; for the transform codec, 2.7-, 11-, 21-, and 40-ms segments could be used; for the CELP codec the segmentation was fixed at 21 ms. Overlapping windows were used for segmentation, but the resulting interactions between the windows were not taken into account in the optimization.

Table 1. Excerpts used in broadcasting test.

| Category | Description |
|---|---|
| Speech with noise (basket) | Commentary of a basket ball match with applause and people shouting |
| Male speech | German male speech |
| Female speech | French female speech |
| Music jazz | Jazz music with a female singer in English |
| Music pop | Pop music with the female singer Tracy Chapman (English) |

Table 2. Excerpts used in storage test.

| Category | Description |
|---|---|
| Music Spanish | Spanish music without any lyrics |
| Male speech | German male speech |
| Harmonic signal | Harpsichord |
| Transient signal | Castanets |
| Music jazz | Jazz music with a female singer in English |
| Music pop | Pop music with the female singer Tracy Chapman (English) |

The individual codecs were operated in a relatively simple mono configuration excluding some of the advanced techniques that have been discussed. This was done to reduce computational complexity and because some of the techniques were not mature enough to operate in a complete coder framework. Specifically the aspects that were incorporated in the coding techniques are described next.

The sinusoidal codec employed a psychoacoustic matching pursuit with constant-amplitude basis functions. The codec used frequency differential linking of parameters, which are RD optimally quantized using point densities as defined in Eqs. (8)–(10). For this codec the resulting quantization errors are neither taken into account in the residue that is given to the following coder nor in the distortion calculations, because the distortion measures were not designed for these errors. So for the cascade of a sinusoidal and a transform codec, the transform coder works on the unquantized residue of the sinusoidal coder. This seems to contradict the conclusions of [27], [25]. There, two approaches are studied, referred to as the parallel and the sequential approach. In the parallel approach a residual for the next coding technique is obtained by subtracting an unquantized reconstruction of the previous technique, whereas in the sequential approach the residue is obtained by subtracting a quantized reconstruction. These papers show that when combining a sinusoidal coder with a waveform coder, a sequential approach leads to higher performance. However, in our study the perceptual distortion measure was not designed for sinusoidal quantization errors, and it was therefore better to use the unquantized residual in this case. With respect to the other coding techniques, the quantized residue was used.

The transform codec used an MDCT filterbank where the prototypical filter shape and filter length were adapted according to the segmentation. Low-pass filtering was applied with variable cutoff frequency. The transform coefficients were quantized according to the description of Eq. (11) and further on.

The CELP codec operated with variable bit rate, where the bit rate was distributed optimally across segments keeping the total bit rate within a frame constant.

In addition to the encoders described, the excitation pattern was encoded at a bit rate of about 4 kbit/s, independent of the RD optimization. At the decoder side this excitation pattern was used to determine a masking curve that was used by the other codecs as common information to efficiently decode the signal. In addition, based on the excitation pattern, a residual noise signal was determined to substitute spectrotemporal components that were not encoded by the other codecs.

### 2.4 Test Results

For the broadcasting scenario the coder has been compared at 20 kbit/s to two codecs: Nero HE-AAC [7] and 3GPP AMR-WB+ [8]. Fig. 9 shows the results. Our coder performs better than HE-AAC (30 points on the 100-point quality scale), which is probably also due to the dominance of speech-oriented signals in the test set. Our coder performs 10 points worse than AMR-WB+, presumably because the bandwidth extension method in our coder was not as mature as that of AMR-WB+ [8]. Furthermore the hybrid coder (CELP+Transf[3]) performs better than the individual coding techniques. An analysis of the encoding process for the individual excerpts confirms our expectations with respect to bit allocation over the coding techniques. Table 3 shows that the RD mechanism predominantly selects the transform coder for the music excerpts, whereas the CELP coder is selected for the speech excerpts. This shows that the hybrid operational RD optimization approach is promising for obtaining increased performance compared to separate coding techniques by effectively exploiting the strengths of the underlying coding techniques.

The coder also benefits from its ability to adapt to the bit rate dynamically, as was demonstrated by the ECQ test. Besides the expected latency and smoothing effects in the judgments, this test showed that the switching between bit rates and the corresponding quality levels does not lead to a quality penalty because the overall quality obtained with dynamic adaptation is similar to that obtained with constant equal-average bit rate (see also [93]). The benefit is therefore clear: dynamic bit-rate adaptation performs significantly better than operation at the lowest (that is, guaranteed) bit rate. Thus the coder can be used for streaming/

[3]CELP+Transf indicates the hybrid coder configuration of the CELP coder followed by the transform coder.

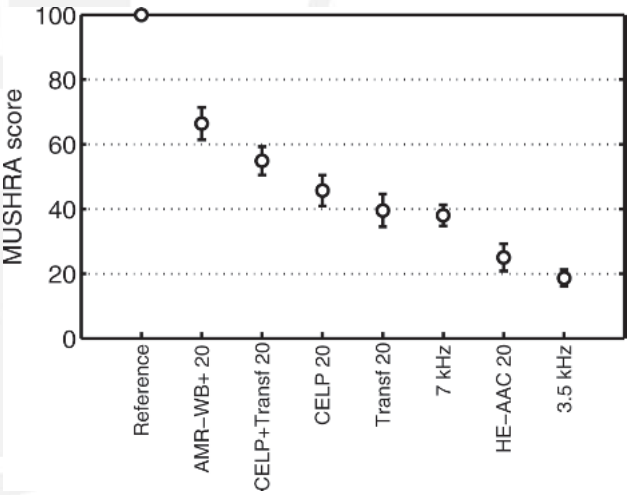



Fig. 9. Results of broadcasting test (20 kbit/s). Error bars denote 95% confidence intervals for mean (15 listeners).

Table 3. Bit-rate distributions in hybrid codec for broadcasting test.*

| | CELP (kbit/s) | Transform (kbit/s) | Excitation Pattern (kbit/s) |
|---|---|---|---|
| Speech with noise (basket) | 4.9 | 12.2 | 4.9 |
| Male speech | 12.0 | 5.0 | 5.0 |
| Female speech | 12.8 | 3.9 | 4.2 |
| Music jazz | 1.0 | 15.3 | 3.6 |
| Music pop | 1.8 | 14.8 | 5.3 |

* Bit rates are expressed as average bit rates across full excerpt. Target bit rate was 20 kbit/s.

transmission and can adapt to the momentary available network bandwidth (such as GSM/GPRS access, shared Internet connection) without a scalability loss.

For the storage scenario the performance of the coder has been evaluated at three bit rates (a high bit rate of 64 kbit/s, an intermediate bit rate of 48 kbit/s, and a low bit rate of 24 kbit/s) and has been compared to the state-of-the-art coders MPEG-2 AAC [1] at the high and intermediate bit rates and MPEG-4 SSC [4] at the low bit rate. For the high bit rate the results are shown in Fig. 10. At this rate the coder is close to transparent, although not as good as the AAC coder, as can be seen in the mean results of the figure. This may be due to noise that was added by the excitation pattern coder at places where it was not desired. At these very high bit rates the relatively coarse quantization of the excitation pattern probably needs to be adapted for a more accurate noise synthesis. It is expected that as a result of such an adaptation, no (or hardly any) noise will be added at this high encoding bit rate. Furthermore, as described in Section 1.5.2, the Huffman tables used for the transform coding technique are optimized for very low and low bit rates. Switchable Huffman tables optimized for different target bit rates, as used, for example, in AAC, may also contribute to a better quality at higher bit rates.

Results for intermediate and low bit rates are shown in Fig. 11 for both the individual coding techniques and our hybrid coder. In addition state-of-the-art coders are shown. As can be seen, at intermediate bit rates (48 kbit/s) the hybrid coder performs better than AAC. This figure also shows the results for the low bit rate (24 kbit/s) at which the coder performs equal to SSC. Also in this test the hybrid combination of different coding techniques leads to improved efficiency compared to using coding techniques alone, again showing that the hybrid RD optimization method allows for the creation of a combined codec that is better than its constituent codecs. It should be noted, however, that at 48 kbit/s the transform coder plays a dominant role and the addition of a sinusoidal coder does not lead to improved performance because at this rate the quality of the hybrid coder and that of the transform coder alone are statistically indistinguishable.

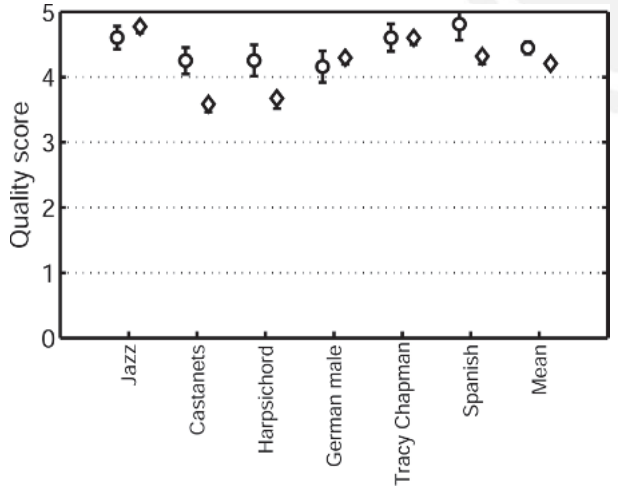



Fig. 10. Results of storage test at 64 kbit/s. Listening test results averaged across 15 listeners for AAC (circles) and our coder (diamonds). For explanation error bars, see Fig. 9.

Tables 4 and 5 show the bit-rate distributions for the target bit rates of 24 kbit/s and 48 kbit/s, respectively. As can be seen, most of the bit rate is given to the transform codec, which is most pronounced for the least tonal signal, the castanet signal. For this transient excerpt the bit budget is almost completely allocated to the transform codec. At 24 kbit/s only 12% of the bit rate available for the hybrid optimized codecs is allocated to the sinusoidal codec. For the remaining excerpts this percentage is 35%. At 48 kbit/s these numbers are 8% and 22%, respectively. So as in [28], and in agreement with Fig. 11, we can conclude from

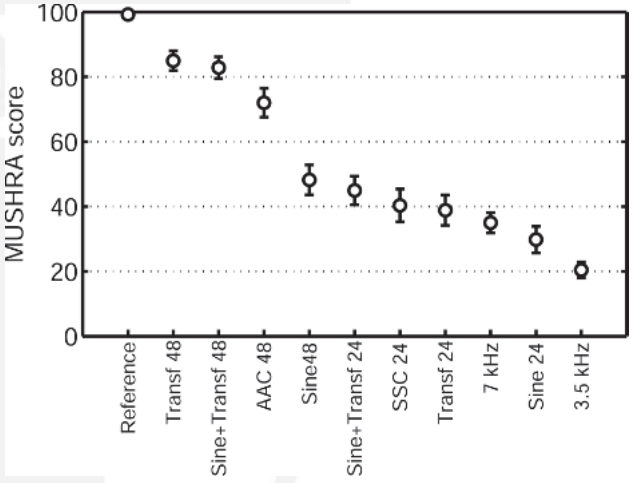



Fig. 11. Results of storage test at 48 and 24 kbit/s (16 listeners). For explanation error bars, see Fig. 9.

Table 4. Bit-rate distributions in hybrid codec for low-storage-capacity test.*

| | Sinusoidal (kbit/s) | Transform (kbit/s) | Excitation Pattern (kbit/s) |
|---|---|---|---|
| Music Spanish | 8.9 | 12.0 | 3.8 |
| Male speech | 5.6 | 15.6 | 5.0 |
| Harmonic signal | 6.2 | 14.9 | 2.9 |
| Transient signal | 2.4 | 17.8 | 3.3 |
| Music jazz | 7.6 | 13.7 | 3.6 |
| Music pop | 8.9 | 12.0 | 5.3 |

* Bit rates are expressed as average bit rates across full excerpt. Target bit rate was 24 kbit/s.

Table 5. Bit-rate distributions in hybrid codec for intermediate-storage-capacity test.*

| | Sinusoidal (kbit/s) | Transform (kbit/s) | Excitation Pattern (kbit/s) |
|---|---|---|---|
| Music Spanish | 8.3 | 38.7 | 3.8 |
| Male speech | 9.8 | 35.0 | 5.0 |
| Harmonic signal | 12.0 | 32.4 | 2.9 |
| Transient signal | 3.5 | 39.9 | 3.3 |
| Music jazz | 9.8 | 35.0 | 3.6 |
| Music pop | 11.2 | 33.3 | 5.3 |

* Bit rates are expressed as average bit rates across full excerpt. Target bit rate was 48 kbit/s.

these numbers that with increasing bit rate a larger percentage of the available rate is given to the transform codec, although in absolute terms the rate spent on the sinusoidal codec tends to increase.

Results for the communication scenario are shown in Fig. 12. The performance of the coder has been evaluated at two delays, a high delay of about 384 ms (Hd) and medium delay of about 43 ms (Md). This has been done for two bit rates, 48 and 20 kbit/s. At 48 kbit/s the sound quality does not suffer much from the more severe delay constraint, but at 20 kbit/s the difference in quality between high and medium delay configurations is rather high. This may have resulted from increased switching between coding techniques because of shorter optimization frames. It should be noted that Fig. 12 presents only results for listening quality, not counting for conversational quality; lower delay will improve conversational quality.

In summary these tests validate that the operational RD optimized hybrid coder can be used in a wide range of applications, in which it performs comparable to the state-of-the-art: the coder performs on average as well as AAC (64 and 48 kbit/s) and SSC (24 kbit/s), better than HE-AAC (20 kbit/s), and slightly worse than AMR-WB+ (20 kbit/s). In addition the results show that the hybrid coder provides an improvement over the individual coding techniques used in this coder. Apparently the RD control was able to select the most efficient coding technique, or combination thereof—even with a coarse bit-distribution step of 25%—exploiting the individual strengths of the different coders effectively.

## 3 DISCUSSION AND CONCLUSION

As also shown by the recent MPEG call for proposals on "Unified Speech and Audio Coding," there is a clear need for generic sound coding technology, which can be used in a wide variety of applications, in contrast to the multitude of existing codecs that are dedicated to a particular application. Operational RD optimization is a promising method for generic sound coding, and this study describes a framework based on these principles. The encoding consists of a control unit, a perceptual distortion measure, and several coding techniques, namely, parametric (sinusoidal) coding, CELP coding, and transform coding, each having its own strengths. Given a particular input signal and constraints such as bit rate and delay, the control unit combines the strengths of the techniques in an efficient way. This approach is general and flexible. For example, the set of coding techniques is not limited to the ones mentioned, but in principle any method could be plugged into the control unit.

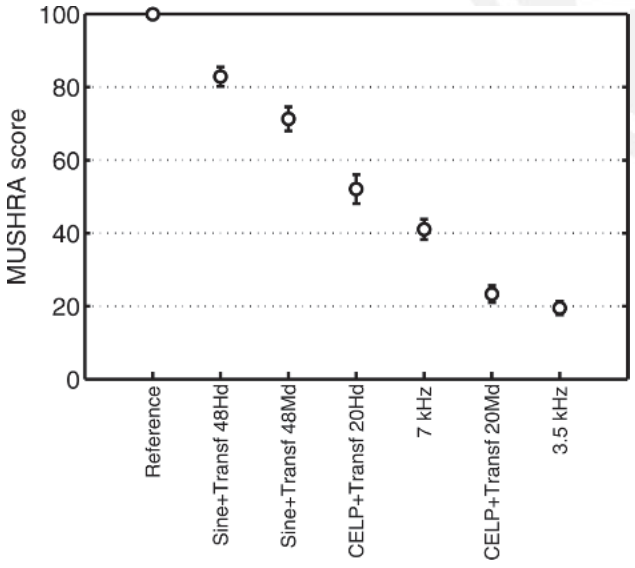



Fig. 12. Results of communication test (48 and 20 kbit/s, 16 listeners). For explanation error bars, see Fig. 9.

The framework has been validated in subjective listening tests. These tests show that the framework can indeed be used in a wide range of applications, such as broadcasting, storage, or communication, in which it is competitive with the state of the art. Furthermore the hybrid framework performs better than the individual coding techniques. In other words, the framework is able to combine the strengths of different coding techniques, which results in higher coding efficiency and flexibility. The framework not only adapts to the input signal, but also to constraints such as bit rate and delay, without the need for specific tuning, as is the case for the state-of-the-art codecs.

It should be noted, however, that more research and development is needed before this technology can be used in applications and services. This especially applies to its complexity, which has to be decreased substantially, and for this the property vector, or similar approaches, can play an important role. Furthermore there is still much room for performance improvement, for example, by improving the coding techniques and especially the distortion measure that determines the bit distribution over these techniques.

In conclusion, this paper shows the feasibility of generic sound coding. A generic sound coding framework has been developed, which obtains its versatility by operational RD optimization techniques. This makes such technology or derivatives thereof candidates for applications of generic sound coding, which will undoubtedly arise in the future.

## 4 ACKNOWLEDGMENT

This work has been funded by E.U. grant IST-2001-34095 (IST program, fifth framework: ARDOR project). The authors would like to thank all contributors to the ARDOR project, especially Aad Rijnberg for ensuring that all individual parts came together in the ARDOR framework and Laetitia Gros for organizing and carrying out part of the listening tests.

## 5 REFERENCES

[1] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc.,* vol. 45, pp. 789–814, (1997 Oct.).

[2] T. Quatieri and R. McAulay, "Audio Signal Processing Based on Sinusoidal Analysis/Synthesis," in *Applications of digital signal processing to audio and acoustics,* M. Kahrs and K. Brandenburg, Eds. (Kluwer Boston, Dordrecht, London, 1998), ch. 9, pp. 343–416.

[3] H. Purnhagen and N. Meine, "HILN—The MPEG-4 Parametric Audio Coding Tools," in *Proc. IEEE Int. Symp. on Circuits and Systems* (Geneva, Switzerland, 2000 May).

[4] ISO/IEC 14496-3:2001/AMD 2:2004, "Information Technology—Coding of Audio-Visual Objects—Part 3: Audio, Amendment 2: Parametric Coding of High Quality Audio," (2004 July).

[5] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Järvinen, "The Adaptive Multi-Rate Wideband Speech Codec (AMR-WB)," *IEEE Trans. Speech Audio Process., special issue,* vol. 10, pp. 620–636 (2002 Nov.).

[6] K. Brandenburg, O. Kunz, and A. Sugiyama, "MPEG-4 Natural Audio Coding," *Signal Process.: Image Commun.,* vol. 15, no. 4–5, pp. 423–444 (2000).

[7] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, "A Closer Look into MPEG-4 High Efficiency AAC," presented at the 115th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 51, p. 1221 (2003 Dec.), convention paper 5871.

[8] J. Mäkinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: A New Audio Coding Standard for 3rd Generation Mobile Audio Services," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 2 (Philadelphia, PA, 2005), pp. 1109–1112.

[9] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, S. Bruhn, and A. Taleb, "Extended AMR-WB for High-Quality Audio on Mobile Devices," *IEEE Commun. Mag.,* pp. 90–97 (2006 May).

[10] S. Ragot, B. Kövesi, R. Trilling, D. Virette, N. Duc, D. Massaloux, S. Proust, B. Geiser, M. Gartner, S. Schandl, H. Taddei, Y. Gao, E. Shlomot, H. Ehara, K. Yoshida, T. Vaillancourt, R. Salami, M. S. Lee, and D. Y. Kim, "ITU-T G.729.1: An 8–32 kbit/s Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice over IP," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Honolulu, HI, 2007).

[11] B. Grill, "A Bit Rate Scalable Perceptual Coder for MPEG-4 Audio," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 45, p. 1005 (1997 Nov.), preprint 4620.

[12] S. A. Ramprashad, "The Multimode Transform Predictive Coding Paradigm," *IEEE Trans. Speech Audio Process.,* vol. 11, pp. 117–129 (2003 Mar.).

[13] F. Riera-Palou, A. C. den Brinker, and A. J. Gerrits, "A Hybrid Parametric-Waveform Approach to Bit Stream Scalable Audio Coding," in *Proc. 38th Asilomar Conf. on Signals, Systems, and Computers* (Pacific Grove, CA, 2004), pp. 2250–2254.

[14] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Sys. Tech. J.,* vol. 27, pp. 379–423 (1948 July).

[15] Y. Shoham and A. Gersho, "Efficient Bit Allocation for an Arbitrary Set of Quantizers," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 36, pp. 1445–1453 (1988 Sept.).

[16] K. Ramchandran and M. Vetterli, "Best Wavelet Packet Bases in a Rate-Distortion Sense," *IEEE Trans. Image Process.,* vol. 2, pp. 160–175 (1993 Apr.).

[17] P. Prandoni, M. Goodwin, and M. Vetterli, "Optimal Time Segmentation for Signal Modeling and Compression," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 3 (Munich, Germany, 1997), pp. 2029–2032.

[18] P. Prandoni, "Optimal Segmentation Techniques for Piecewise Stationary Signals," Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, Switzerland (1999).

[19] P. Prandoni and M. Vetterli, "R/D Optimal Linear Prediction," *IEEE Trans. Speech Audio Process.,* vol. 8, pp. 646–655 (2000 Nov.).

[20] R. Heusdens and S. van de Par, "Rate-Distortion Optimal Sinusoidal Modeling of Audio and Speech Using Psychoacoustical Matching Pursuits," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 2, (Orlando, FL, 2002), pp. 1809–1812.

[21] S. van de Par and A. Kohlrausch, "Application of a Spectrally Integrating Auditory Filterbank Model to Audio Coding," in *Fortschritte der Akustik,* Plenarvorträge der 28. Deutschen Jahrestagung für Akustik (DAGA-02) (Bochum, Germany, 2002).

[22] C. Bauer and M. Vinton, "Joint Optimization of Scale Factors and Huffman Code Books for MPEG-4 AAC," *IEEE Trans. Signal Process.,* vol. 54, pp. 177–189 (2006 Jan.).

[23] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding* (Wiley, Hoboken, NJ, 2007).

[24] B. C. J. Moore, *An Introduction to the Psychology of Hearing* (Academic Press, London, 1998).

[25] R. Vafin and W. B. Kleijn, "Rate-Distortion Optimized Quantization in Multistage Audio Coding," *IEEE Trans. Audio, Speech, Language Process.,* vol. 14, pp. 311–320 (2006 Jan.).

[26] B. Kövesi, D. Massaloux, D. Virette, and J. Bensa, "Integration of a CELP Coder in the ARDOR Universal Sound Codec," in *Proc. Interspeech 2006* (Pittsburgh, PA, 2006).

[27] R. Vafin and W. B. Kleijn, "Towards Optimal Quantization in Multistage Audio Coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 4 (Montreal, Canada, 2004), pp. 205–208.

[28] N. H. van Schijndel and S. van de Par, "Rate-Distortion Optimized Hybrid Sound Coding," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, 2005), pp. 235–238.

[29] Z. Xiong, C. Herley, K. Ramchandran, and M. T. Orchard, "Flexible Time Segmentations for Time-Varying Wavelet Packets," in *Proc. IEEE Int. Symp. on Time–Frequency and Time-Scale Analysis* (1994), pp. 9–12.

[30] Z. Xiong, K. Ramchandran, C. Herley, and M. T. Orchard, "Flexible Tree-Structured Signal Expansions Using Time-Varying Wavelet Packets," *IEEE Trans. Signal Process.,* vol. 45, pp. 333–345 (1997 Feb.).

[31] R. Bellman, *Dynamic Programming* (Princeton University Press, Princeton, NJ, 1957).

[32] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models* (Prentice-Hall, Englewood Cliffs, NJ, 1987).

[33] C. A. Rødbro, J. Jensen, and R. Heusdens, "Adaptive Time-Segmentation for Speech Coding with Limited Time Delay," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. I. (Montreal, Canada, 2004), pp. 465–468.

[34] N. H. van Schijndel and G. d'Ambrosio, "On Delay in Parametric Audio Coding with Adaptive Segmentation," in *Proc. 1st IEEE BENELUX/DSP Valley Signal Processing Symposium* (*SPS-DARTS'05*) (Antwerp, Belgium, 2005), pp. 37–40.

[35] O. A. Niamut and R. Heusdens, "RD Optimal Time Segmentations for the Time-Varying MDCT," in *Proc. XII European Signal Processing Conf.* (*EUSIPCO*) (Vienna, Austria, 2004), pp. 1649–1652.

[36] O. A. Niamut and R. Heusdens, "Optimal Time Segmentation for Overlap–Add Systems with Variable Amount of Window Overlap," *IEEE Signal Process. Lett.,* vol. 12, pp. 665–668 (2005 Oct.).

[37] O. A. Niamut and R. Heusdens, "Flexible Frequency Decompositions for Cosine-Modulated Filter Banks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 5 (Hong Kong, China, 2003), pp. 449–452.

[38] O. A. Niamut and R. Heusdens, "Subband Merging in Cosine-Modulated Filter Banks," *IEEE Signal Process. Lett.,* vol. 10, pp. 111–114 (2003 Apr.).

[39] O. A. Niamut and R. Heusdens, "RD Optimal Temporal Noise Shaping for Transform Audio Coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Toulouse, France, May 2006), pp. 189–192.

[40] J. Herre and J. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping," presented at the 101st Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 44, p. 1175 (1996 Dec.), preprint 4384.

[41] S. Buus, E. Schorer, M. Florentine, and E. Zwicker, "Decision Rules in Detection of Simple and Complex Tones," *J. Acoust. Soc. Am.,* vol. 80, pp. 1646–1657 (1986).

[42] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A Perceptual Model for Sinusoidal Audio Coding Based on Spectral Integration," *EURASIP J. Appl. Signal Process.,* special issue on Anthropomorphic Signal Processing, pp. 1292–1304 (2005 June).

[43] T. Dau, D. Püschel, and A. Kohlrausch, "A Quantitative Model of the 'Effective' Signal Processing in the Auditory System. I. Model Structure," *J. Acoust. Soc. Am.,* vol. 99, pp. 3615–3622 (1996).

[44] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal Modeling Using Psychoacoustic-Adaptive Matching Pursuits," *IEEE Signal Process. Lett.,* vol. 9, pp. 262–265 (2002 Aug.).

[45] R. Heusdens, J. Jensen, W. B. Kleijn, V. Kot, O. A. Niamut, S. van de Par, N. H. van Schijndel, and R. Vafin, "Bit-Rate Scalable Intraframe Sinusoidal Audio Coding Based on Rate-Distortion Optimization," *J. Audio Eng. Soc.,* vol. 54, pp. 167–188 (2006 Mar.).

[46] M. G. Christensen and S. H. Jensen, "On Perceptual Distortion Minimization and Nonlinear Least-Squares Frequency Estimation," *IEEE Trans. Audio, Speech, Language Process.,* vol. 14, pp. 99–109 (2006 Jan.).

[47] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Amplitude Modulated Sinusoidal Models for Audio Modeling and Coding," in *Knowledge-Based Intelligent Information and Engineering Systems,* Lecture Notes in Artificial Intelligence, V. Palade, R. J. Howlett, and L. C. Jain, Eds. (Springer, New York, 2003), vol. 2773, pp. 1334–1342.

[48] M. G. Christensen, S. van de Par, S. H. Jensen, and S. V. Andersen, "Multiband Amplitude Modulated Sinusoidal Audio Modeling," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 4 (2004), pp. 169–172.

[49] M. G. Christensen, A. Jakobsson, S. V. Andersen, and S. H. Jensen, "Amplitude Modulated Sinusoidal Signal Decomposition for Audio," *IEEE Signal Process. Lett.,* vol. 13, pp. 389–392 (2006 July).

[50] M. G. Christensen and S. van de Par, "Efficient Parametric Coding of Transients," *IEEE Trans. Audio, Speech, Language Process.,* vol. 14, pp. 1340–1351 (2006 July).

[51] M. G. Christensen and S. H. Jensen, "Computationally Efficient Amplitude Modulated Sinusoidal Audio Coding Using Frequency-Domain Linear Prediction," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 5 (2006), pp. 61–64.

[52] M. G. Christensen and S. H. Jensen, "New Results in Rate-Distortion Optimized Parametric Audio Coding," presented at the 120th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 54, p. 723 (2006 July/Aug.), convention paper 6808.

[53] R. Vafin and W. B. Kleijn, "Entropy-Constrained Polar Quantization: Theory and an Application to Audio Coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 2 (Orlando, FL, 2002), pp. 1837–1840.

[54] R. Vafin and W. B. Kleijn, "Entropy-Constrained Polar Quantization and Its Application to Audio Coding," *IEEE Trans. Speech Audio Process.,* vol. 13, pp. 220–232, (2005 Mar.).

[55] R. Vafin, D. Prakash, and W. B. Kleijn, "On Frequency Quantization in Sinusoidal Audio Coding," *IEEE Signal Process. Lett.,* vol. 12, pp. 210–213, (2005 Mar.).

[56] R. Vafin and W. B. Kleijn, "Jointly Optimal Quantization of Parameters in Sinusoidal Audio Coding," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, 2005), pp. 247–250.

[57] P. Korten, J. Jensen, and R. Heusdens, "High Resolution Spherical Quantization of Sinusoidal Parameters Using a Perceptual Distortion Measure," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 3 (Philadelphia, PA, 2005), pp. 177–180.

[58] P. Korten, J. Jensen, and R. Heusdens, "High-Resolution Spherical Quantization of Sinusoidal Parameters," *IEEE Trans. Audio, Speech, Language Process.,* vol. 15, pp. 966–981 (2007 Mar.).

[59] R. Heusdens and J. Jensen, "Jointly Optimal Time Segmentation, Component Selection and Quantization for

Sinusoidal Coding of Audio and Speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 3 (Philadelphia, PA, 2005), pp. 193–196.

[60] R. Heusdens, J. Jensen, and P. Korten, "Jointly Optimal Time Segmentation, Distribution, and Quantization for Sinusoidal Coding of Audio and Speech," presented at the 119th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, p. 1227 (2005 Dec.), convention paper 6596.

[61] R. Heusdens, J. Jensen, P. Korten, and R. Vafin, "Rate-Distortion Optimal High-Resolution Differential Quantisation for Sinusoidal Coding of Audio and Speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, 2005), pp. 243–246.

[62] M. H. Larsen, M. G. Christensen, and S. H. Jensen, "Variable Dimension Trellis-Coded Quantization of Sinusoidal Parameters," *IEEE Signal Process. Lett.,* vol. 15, pp. 17–20 (2008).

[63] M. H. Larsen, M. G. Christensen, and S. H. Jensen, "Multiple Description Trellis-Coded Quantization of Sinusoidal Parameters," *IEEE Trans. Signal Process.,* vol. 56, pt. 2, pp. 5287–5291 (2008 Oct.).

[64] O. Wübbolt, "*Codierung von Erregungsmustern zur Steuerung einer transformationsbasierten Audiocodierung*" (*in German*), Ph.D. thesis, Universität Hannover, Fortschritt-Berichte VDI, Reihe 10, no. 775 (2006).

[65] N. Meine and B. Edler, "Improved Quantization and Lossless Coding for Subband Audio Coding," presented at the 118th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, p. 699 (2005 July/Aug.), convention paper 6468.

[66] O. Niemeyer and B. Edler, "Detection and Extraction of Transients for Audio Coding," presented at the 120th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 54, p. 723 (2006 July/Aug.), convention paper 6811.

[67] 3GPP TS 26.190, "Speech Codec Speech Processing Functions; Adaptive Multi-Rate-Wideband (AMR-WB) Speech Codec; Transcoding Functions," http://www.3gpp.org (2004).

[68] S. van de Par, V. Kot, and N. H. van Schijndel, "Scalable Noise Coder for Parametric Sound Coding," presented at the 118th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, p. 699 (2005 July/Aug.), convention paper 6465.

[69] T. V. Sreenivas and M. Dietz, "Improved AAC Performance @ <64kb/s Using VQ," presented at the 104th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 46, p. 581 (1998 June), preprint 4750.

[70] W. A. Pearlman, A. Islam, N. Nagaraj, and A. Said, "Efficient, Low-Complexity Image Coding with a Set-Partitioning Embedded Block Coder," *IEEE Trans. Circuits Sys. for Video Technol.,* vol. 14, pp. 1219–1235 (2004 Nov.).

[71] O. Niemeyer and B. Edler, "Efficient Coding of Excitation Patterns Combined with a Transform Audio Coder," presented at the 118th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, p. 699 (2005 July/Aug.), convention paper 6466.

[72] D. A. Huffman, "A Method for the Construction of Minimum-Redundancy Codes," *Proc. IRE,* vol. 40, pp. 1098–1101 (1952 Sept.).

[73] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC–Analysis/Synthesis Audio Codec for Very Low Bit Rates," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 44, p. 636 (1996 July/Aug.), preprint 4179.

[74] K. Hamdy, M. Ali, and A. Tewfik, "Low Bit-Rate High-Quality Audio Coding with Combined Harmonic and Wavelet Representation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Atlanta, GA, 1996), pp. 1045–1048.

[75] J. Jensen and R. Heusdens, "Schemes for Optimal Frequency-Differential Encoding of Sinusoidal Model Parameters," *Signal Process.,* vol. 83, pp. 1721–1735 (2003 Aug.).

[76] J. Jensen and R. Heusdens, "A Comparison of Differential Schemes for Low-Rate Sinusoidal Audio Coding," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, 2003), pp. 205–208.

[77] J. Lindblom, J. Plasberg, and R. Vafin, "Flexible Sum–Difference Stereo Coding Based on Time-Aligned Signal Components," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, 2005), pp. 255–258.

[78] J. Johnston and A. Ferreira, "Sum–Difference Stereo Transform," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 2 (1992), pp. 569–572.

[79] F. Baumgarte and C. Faller, "Binaural Cue Coding—Part I: Psychoacoustic Fundamentals and Design Principles," *IEEE Trans. Speech Audio Process.,* vol. 11, pp. 509–519 (2003).

[80] C. Faller and F. Baumgarte, "Binaural Cue Coding—Part II: Schemes and Applications," *IEEE Trans. Speech Audio Process.,* vol. 11, pp. 520–531 (2003).

[81] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "High-Quality Parametric Spatial Audio Coding at Low Bitrates," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 52, p. 800 (2004 July/Aug.), convention paper 6072.

[82] F. Nordén, S. V. Andersen, S. H. Jensen, and W. B. Kleijn, "Property Vector Based Distortion Estimation," in *Proc. 38th Annual Conf. on Signals, Systems, and Computers* (Pacific Grove, CA, 2004).

[83] F. Nordén, M. G. Christensen, and S. H. Jensen, "Open Loop Rate-Distortion Optimized Audio Coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 3, (Philadelphia, PA, 2005), pp. 161–164.

[84] C. A. Rødbro, M. G. Christensen, F. Nordén, and S. H. Jensen, "Low Complexity Rate-Distortion Optimized Time-Segmentation for Audio Coding," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, 2005), pp. 231–234.

[85] W. R. Gardner and B. D. Rao, "Theoretical Analysis of the High-Rate Vector Quantization of LPC Parameters," *IEEE Trans. Speech Audio Process.,* vol. 3, pp. 367–381 (1995 Sept.).

[86] J. H. Plasberg, D. Y. Zhao, and W. B. Kleijn, "The Sensitivity Matrix for a Spectro-Temporal Auditory Model," in *Proc. XII European Signal Processing Conf.* (*EUSIPCO*) (Vienna, Austria, 2004), pp. 1673–1676.

[87] J. H. Plasberg and W. B. Kleijn, "The Sensitivity Matrix: Using Advanced Auditory Models in Speech and Audio Processing," *IEEE Trans. Audio, Speech, Language Process.,* vol. 15, pp. 310–319 (2007 Jan.).

[88] J. Johnston, "Estimation of Perceptual Entropy Using Noise Masking Criteria," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (New York, 1988), pp. 2524–2527.

[89] O. A. Niamut, R. Heusdens, and H. J. Lincklaen Arriëns, "Upfront Time Segmentation Methods for Transform Coding of Audio," presented at the 119th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, pp. 1222, 1223 (2005 Dec.), convention paper 6585.

[90] ITU-R BS.1534, "Method for the Subjective Assessment of Intermediate Quality Level Coding Systems," International Telecommunications Union, Geneva, Switzerland (2001), http://www.itu.int.

[91] ITU-T P.880, "Continuous Evaluation of Time-Varying Speech Quality," International Telecommunications Union, Geneva, Switzerland (2004), http://www.itu.int.

[92] ITU-R BS.1116-1, "Method for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (1997), http://www.itu.int.

[93] N. H. van Schijndel, L. Gros, and S. van de Par, "Dynamic Bit-Rate Adaptation for Speech and Audio," presented at the 123rd Convention of the Audio Engineering Society, (*Abstracts*) http://www.aes.org/events/123/123rdWrapUp.pdf, (2007 Oct.), convention paper 7288.

## THE AUTHORS

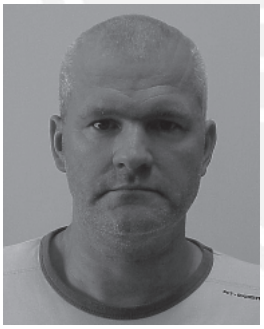

N. H. van Schijndel


J. Bensa


M. G. Christensen


C. Colomes


B. Edler


R. Heusdens


J. Jensen


S. H. Jensen


W. B. Kleijn


V. Kot


B. Kövesi


J. Lindblom


D. Massaloux


O. A. Niamut


F. Nordén


J. H. Plasberg


R. Vafin


S. van de Par


D. Virette


O. Wübbolt

Nicole H. van Schijndel received an M.Sc. degree in physics from Radboud University Nijmegen, The Netherlands, in 1995, and a Ph.D. degree from VU University Amsterdam, The Netherlands, in 2000.

In 1999 she joined the Digital Signal Processing Group at Philips Research Laboratories, Eindhoven, The Netherlands. Her main research activities are in the fields of audio coding, speech enhancement, speech perception, and hearing impairment. She was the international project leader of the E.U.-funded ARDOR project.

●

Julien Bensa obtained a Master's degree in acoustics, signal processing, and informatics applied to music from the Pierre et Marie Curie University, Paris, France, in 1998. In 2001 he worked at the Center for Computer Research in Music and Acoustics (CCRMA) at Stanford University, Stanford, CA. He received a Ph.D. degree in acoustics and signal processing from the University of Aix-Marseille II, France, in 2003 for his work on the analysis and synthesis of piano sounds using physical and signal models.

In 2004 he worked at the Laboratoire d'Acoustique Musicale, Paris, on the perceived quality of synthesis musical tones. He joined the ARDOR project at France Telecom, Lannion, in 2005 and participated in the integration of the CELP coder in the ARDOR framework.

●

Mads Græsbøll Christensen was born in Copenhagen, Denmark, in March 1977. He received M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University, Aalborg, Denmark, where he is currently an assistant professor in the Department of Electronic Systems. He has been a visiting researcher at Philips Research Labs, the Ecole Nationale Supérieure des Télécommunications (ENST), and Columbia University.

Dr. Christensen has received several awards, namely, an IEEE International Conference on Acoustics, Speech, and Signal Processing Student Paper Contest Award, the Spar Nord Foundation's Research Prize awarded annually for an excellent Ph.D. thesis, and a Danish Independent Research Council's Young Researcher's Award. His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, and coding of speech and audio signals.

●

Catherine Colomes graduated in electronics and computer science from an electronic engineering high school. After receiving a Master's degree in music technology from the University of York, England, she completed a doctorate thesis on perceptual objective measures applied to low-bit-rate audio codecs.

She then joined France Telecom/Orange labs as a research and development engineer and currently works in the domain of subjective and objective audio quality assessment and real-time implementation.

●

Bernd Edler studied electrical engineering in Erlangen, Germany, and received a Dipl.-Ing. degree in 1985.

From 1986 to 1993 he was a research assistant at the Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, University of Hannover, Germany, where he received a Dr.-Ing. degree in 1995. His major research activities were in the field of filter banks and their applications in audio coding. He contributed to the window switching and aliasing reduction techniques for the development of the filter bank used in the ISO/MPEG-1/2 Layer 3 audio coding standard.

Since 1993 he has been head of the Systems Technologies Division at the Information Technology Laboratory of the University of Hannover. He has actively participated in the development of the MPEG-4 Audio Standard. From 1998 to 1999 he was a research visitor at Bell Laboratories, Murray Hill, NJ, where he developed the prefilter technique used today in ultralow-delay audio coding. His current research fields cover very low bit-rate audio coding, models for auditory perception, and signal processing for cochlear implants.

Richard Heusdens received M.Sc. and Ph.D. degrees from Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively.

In 1992 he joined the Digital Signal Processing Group at the Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997 he joined the Circuits and Systems Group of Delft University of Technology, where he was a postdoctoral researcher. In 2000 he moved to the Information and Communication Theory (ICT) Group, where he became an assistant professor responsible for the audio and speech processing activities within the ICT group. Since 2002 he has been an associate professor in the Department of Mediamatics at Delft University of Technology. He held visiting positions at KTH (Royal Institute of Technology, Sweden) in 2002 and 2008. He is involved in research projects that cover subjects such as audio and speech coding, speech enhancement, and digital watermarking of audio.

●

Jesper Jensen received M.Sc. and Ph.D. degrees in electrical engineering from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2001 he was with the Center for PersonKommunikation (CPK), Aalborg University, as a researcher, Ph.D. student, and assistant research professor. In 1999 he was a visiting researcher at the Center for Spoken Language Research, University of Colorado, Boulder. From 2000 to 2007 he was a postdoctoral researcher and assistant professor at the Delft University of Technology, Delft, The Netherlands. He is currently working at Oticon, Smørum, Denmark. His main research interests are digital speech and audio signal processing, including coding, synthesis, and enhancement.

●

Søren Holdt Jensen received an M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and a Ph.D. degree from the Technical University of Denmark, Lyngby, in 1995.

He has been with the Telecommunications Laboratory, Telecom Denmark, Copenhagen, the Electronics Institute of the Technical University of Denmark, the Scientific Computing Group of the Danish Computing Center for Research and Education (UNI-C), Lyngby, the Electrical Engineering Department, Katholieke Universiteit Leuven, Leuven, Belgium, and the Center for PersonKommunikation (CPK), Aalborg University. He is now a professor in the Department of Electronic Systems at Aalborg University and head of its Multimedia Information and Signal Processing Section. His research activities are in statistical signal processing, communication signal processing, speech and audio processing, and multimedia technology.

Dr. Jensen has been an associate editor for the *IEEE Transactions on Signal Processing* and is currently a member of the editorial board of *EURASIP Journal on Advances in Signal Processing* and an associate editor for *Elsevier Signal Processing*. He was the guest editor on special issues for the *EURASIP Journal on Applied Signal Processing* on anthropomorphic processing of audio and speech and digital signal processing in hearing aids and cochlear implants. He is a recipient of a European Community Marie Curie Fellowship and former chair of the IEEE Denmark Section and its Signal Processing Chapter.

●

W. Bastiaan Kleijn received a Ph.D. degree in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 1991, a Ph.D. degree in soil science and an M.S. degree in physics, both in 1981 from the University of California, and an M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1984. He worked on speech processing at AT&T Bell Laboratories from 1984 to 1996, first in development and from 1990 in research. He moved to KTH in 1996 and has been a visiting professor at Delft University of Technology and Vienna University of Technology, Vienna, Austria. At present he is a professor and head of the Sound and Image Processing Laboratory at the School of Electrical Engineering at KTH (Royal Institute of Technology) in Stockholm, Sweden. He is also a founder and former chair of Global IP Solutions, where he remains chief scientist.

Dr. Kleijn has authored and coauthored over 175 peer-reviewed papers and was awarded 28 U.S. patents. He is a fellow of the IEEE.

●

Valery Kot was born in Minsk, Belarus, in January 1971. He received an M.Sc. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology in 1994.

Until 1999 he was a research scientist at the Institute of Informatics Problems of the Russian Academy of Sciences in Moscow. In 2002 he joined Philips Research Laboratories, Eindhoven, The Netherlands. His interests include digital signal processing theory and its application to audio and video signals.

●

Balázs Kövesi was born in Pécs, Hungary, in 1968. He received a degree in electrical engineering from the Technical University of Budapest in 1992, an M.Sc. degree from Telecom Bretagne, France, in 1993, and a Ph.D. degree from the University of Rennes I, France, in 1997.

He joined the Speech and Audio Coding Group of France Télécom / Orange as a postdoctoral fellow in 1997 and as a research engineer in 1998. His main research interests are connected to speech and audio compression.

●

Jonas Lindblom was born in Skellefteå, Sweden, on March 22, 1974. He received an M.Sc. degree in electrical engineering and a Ph.D. degree in information theory from Chalmers University of Technology, Göteborg, Sweden, in 1998 and 2003, respectively. The title of his Ph.D. thesis is "Coding Speech for Packet Networks."

During 2004–2005 he was a postdoctoral researcher with the Sound and Image Processing Lab at the Royal Institute of Technology (KTH) in Stockholm, working mainly on new semiparametric audio coding methods. He is now with Skype Technologies, Stockholm. His research interests are within the fields of speech, audio and video processing, and transmission.

●

Dominique Massaloux was born in France on October 28, 1956. She graduated from the Ecole Nationale Supérieure des Télécommunications, Paris (Telecom ParisTech), France, in 1979 and received an engineering doctorate in automatics and signal processing at the Image Laboratory of Telecom ParisTech in 1982.

She joined France Télécom Research Center in 1984 and its Speech Coding Department in 1985.

She has been taking part in the standardization of speech coding algorithms in ETSI and ITU-T and was head of the speech coding and signal processing team between 1998 and 2006. She is now in charge of research programs on speech, audio, and video coding and on interface technologies.

Omar A. Niamut received M.Sc. and Ph.D. degrees in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2001 and 2006, respectively.

From 2001 to 2005 he worked on the national and European research projects SiCAS and ARDOR on algorithms for universal audio coding. During his Ph.D. studies at the Information and Communication Theory Group he investigated the use of rate-distortion principles for filter bank-based audio coding. Since 2006 he has been with the Multimedia Technology Department at TNO Information and Communication Technology, Delft, where he is working on projects related to IPTV networks, services, and standardization. His current focus is on the combination of IMS and IPTV to leverage social TV and service blending, the perceived quality of experience of HD, and audiovisual signal processing for digital TV.

●

Fredrik Nordén was born in Mölndal, Sweden, in 1969. He received an M.S. degree in electrical engineering in 1994 and a Ph.D. degree in information theory in 2003, both from Chalmers University of Technology, Göteborg, Sweden.

From 1994 to 1998 he was with the Signal Processing Group at Saab Bofors Dynamics, Göteborg, Sweden. From 1998 to 2003 he was with the Information Theory Group, Chalmers University of Technology. From 2003 to 2004 he worked at the Department of Communication Technology, Aalborg University, Denmark. From 2004 to 2005 he was with the Mobile Devices Department at Teleca Systems AB, where he worked as a consultant in the area of signal processing and software development. From 2005 to 2007 he was with the System Design Department at Imego AB, Sweden, where he performed signal processing for microsensor systems. He is currently with Ericsson AB, Sweden, performing software development for 3G mobile base stations.

●

Jan H. Plasberg received a Dipl.-Ing. degree in electrical engineering from Aachen University of Technology (RWTH), Aachen, Germany, in 2002. He is currently pursuing a Ph.D. degree at the School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden.

From 2000 to 2001 he was a student researcher at the Institute of Communication Systems and Data Processing, RWTH, Aachen. During 2001 he was an intern at Global IP Sound, Stockholm, Sweden. From 2002 to 2008 he was employed as a Ph.D. student at KTH, Stockholm. He is now with Skype Sweden AB, Stockholm, Sweden. His research interests include speech and audio processing, source coding, and Bayesian classification.

Mr. Plasberg received the Young Authors Award of EUSIPCO-2004, Vienna, Austria.

●

Renat Vafin received an engineering degree in telecommunications from Tallinn Technical University, Tallinn, Estonia, in 1994, an M.Sc. degree in digital communication systems and technologies from Chalmers University of Technology, Gothenburg, Sweden, in 1996, and a Ph.D. degree in acoustic signal processing from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2004.

From 1993 to 1998 he was an engineer and teaching assistant with the Department of Radio and Communications Engineering, Tallinn Technical University. From 2004 to 2005 he was a postdoctoral fellow at the Sound and Image Processing Laboratory, Department of Signals, Sensors, and Systems, KTH. He is now with Skype Technologies OÜ, Tallinn, Estonia. His research interests include audio and video processing and source coding.

●

Steven van de Par studied physics at the Eindhoven University of Technology, Eindhoven, The Netherlands, and received a Ph.D. degree in 1998 from the Institute for Perception Research on a topic related to binaural hearing. As a postdoctoral researcher at the same institute he studied auditory–visual interaction.

He was a guest researcher at the University of Connecticut Health Center and at the Center for Applied Hearing Research at the Technical University of Denmark. In 2000 he joined Philips Research Laboratories, Eindhoven, The Netherlands.

Since 2007 Dr. van de Par has been a committee member of the Netherlands Section of the Audio Engineering Society. His work on auditory and multisensory perception, low-bit-rate audio coding, and music information retrieval has resulted in over 100 journal papers, book chapters, and conference publications.

●

David Virette received a Dipl.-Ing. degree in electronics and computer science from the ENSSAT, Lannion, France, in 2000 and an M.Sc. degree in signal processing and telecommunications from the University of Rennes, Rennes, France, in 2000.

Since 2000 he has been with France Télécom/Orange, where he is a research engineer.

Mr. Virette has contributed to MPEG standards and to the development of ITU-T speech and audio coding standards, namely, G.729.1 and G.718. His main research interests include three-dimensional audio and speech/audio coding.

●

Oliver Wübbolt né Niemeyer received Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from the Leibnitz Universität, Hannover, Germany, in 1998 and 2006, respectively. He also received a Dipl.-Wirt. Ing. degree in engineering management from the Fern Universität in Hagen, Germany, in 2006.

In 2005 he joined the Digital Audio Laboratory, Corporate Research, Thomson, Hannover, as a senior development engineer, working on audio coding. His research interests include signal processing, psychoacoustics, speech, audio, and video coding.

-